Lab 05 - Spark Streaming Practice Report

Instructor: PhD. Dao Thanh Chung
Group 7:
- Chu Hoang Duong - 20194429
- Nguyen Van Thanh Tung - 20190090
- Nguyen Vu Thien Trang - 20194459

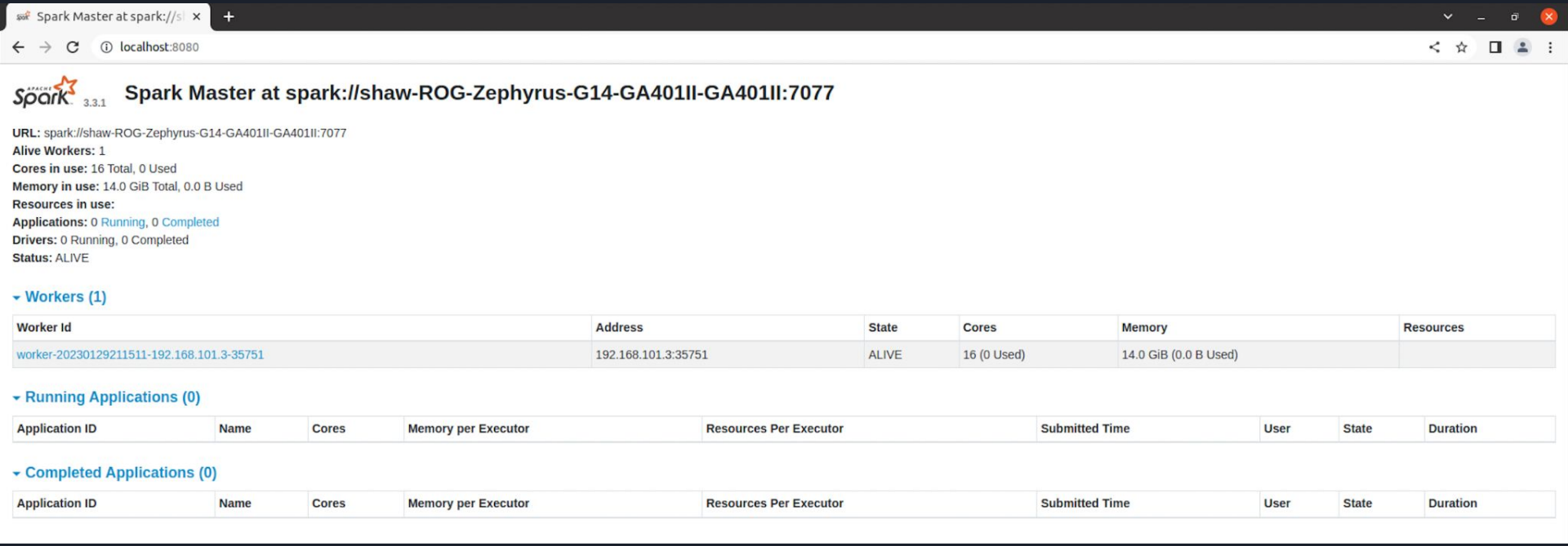
1. Spark Settings

We started Spark Standalone Cluster Mode

```
shaw@shaw-ROG-Zephyrus-G14-GA401II-GA401II:~$ $SPARK_HOME/sbin/start-all.sh
starting org.apache.spark.deploy.master.Master, logging to /home/shaw/spark-practice/spark-3.3.1-bin-hadoop3/logs/spark-shaw-org.apache.spark.deploy.master.Master-1-shaw-ROG-Zephyrus-G14-GA401II-GA401II.out
localhost: starting org.apache.spark.deploy.worker.Worker, logging to /home/shaw/spark-practice/spark-3.3.1-bin-hadoop3/logs/spark-shaw-org.apache.spark.deploy.worker.Worker-1-shaw-ROG-Zephyrus-G14-GA401II-GA401II.out
shaw@shaw-ROG-Zephyrus-G14-GA401II-GA401II:~$ jps
3008 Master
3209 Worker
3307 Jps
shaw@shaw-ROG-Zephyrus-G14-GA401II-GA401II:~$
```

1. Spark Settings

Obtain a cluster with 1 master node and 1 worker



2. Socket Stream

Source Code: SocketStream.scala

SocketStream.scala

```scala
import org.apache.spark._
import org.apache.spark.SparkContext._
import org.apache.spark.streaming.StreamingContext
import org.apache.spark.streaming.StreamingContext._
import org.apache.spark.streaming.dstream.DStream
import org.apache.spark.streaming.Duration
import org.apache.spark.streaming.Seconds

object SocketStream {
  def main(args: Array[String]) {
    val conf = new SparkConf().setAppName("Socket-Stream")

    // Create a StreamingContext with a 1-second batch size from a SparkConf
    val ssc = new StreamingContext(conf, Seconds(1))

    // Create a DStream using data received after connecting to port 7777 on the local machine
    val lines = ssc.socketTextStream("localhost", 7777)

    // Filter our DStream for lines with "error"
    val errorLines = lines.filter(_.contains("error"))

    // Print out the lines with errors
    errorLines.print()

    // Start our streaming context
    ssc.start()

    // Wait for the job to finish
    ssc.awaitTermination()
  }
}
```

Settings: build.sbt

build.sbt

```
name := "socket-stream"

version := "1.0"

scalaVersion := "2.12.15"

libraryDependencies ++= Seq {
  "org.apache.spark" %% "spark-core" % "3.3.1" % "provided";
  "org.apache.spark" %% "spark-streaming" % "3.3.1"
}
```

2. Socket Stream

Package the project by sbt

```
shaw@shaw-ROG-Zephyrus-G14-GA401II-GA401II:~/spark-practice/spark-3.3.1-bin-hadoop3/examples/socket-stream$ find .
.
./build.sbt
./src
./src/main
./src/main/scala
./src/main/scala/SocketStream.scala
shaw@shaw-ROG-Zephyrus-G14-GA401II-GA401II:~/spark-practice/spark-3.3.1-bin-hadoop3/examples/socket-stream$ sbt clean package
[info] Updated file /home/shaw/spark-practice/spark-3.3.1-bin-hadoop3/examples/socket-stream/project/build.properties: set sbt.version to 1.8.2
[info] welcome to sbt 1.8.2 (Ubuntu Java 11.0.17)
[info] loading project definition from /home/shaw/spark-practice/spark-3.3.1-bin-hadoop3/examples/socket-stream/project
[info] loading settings for project socket-stream from build.sbt ...
[info] set current project to socket-stream (in build file:/home/shaw/spark-practice/spark-3.3.1-bin-hadoop3/examples/socket-stream/)
[success] Total time: 0 s, completed Jan 29, 2023, 4:22:53 PM
[info] compiling 1 Scala source to /home/shaw/spark-practice/spark-3.3.1-bin-hadoop3/examples/socket-stream/target/scala-2.12/classes ...
[success] Total time: 5 s, completed Jan 29, 2023, 4:22:58 PM
```

2. Socket Stream

Obtain jar file

- A JAR ("Java archive") is a package file format typically used to aggregate many Java class files and associated metadata resources (text, images, etc.) into one file for distribution

```
./target/streams/compile/managedClasspath/_global/streams
./target/streams/compile/managedClasspath/_global/streams/export
./target/streams/compile/internalDependencyClasspath
./target/streams/compile/internalDependencyClasspath/_global
./target/streams/compile/internalDependencyClasspath/_global/streams
./target/streams/compile/internalDependencyClasspath/_global/streams/export
./target/streams/compile/internalDependencyClasspath/_global/streams/out
./target/streams/compile/_global
./target/streams/compile/_global/_global
./target/streams/compile/_global/_global/discoveredMainClasses
./target/streams/compile/_global/_global/discoveredMainClasses/data
./target/streams/compile/_global/_global/compileOutputs
./target/streams/compile/_global/_global/compileOutputs/previous
./target/streams/compile/bspReporter
./target/streams/compile/bspReporter/_global
./target/streams/compile/bspReporter/_global/streams
./target/streams/compile/bspReporter/_global/streams/out
./target/streams/compile/unmanagedClasspath
./target/streams/compile/unmanagedClasspath/_global
./target/streams/compile/unmanagedClasspath/_global/streams
./target/streams/compile/unmanagedClasspath/_global/streams/export
./target/streams/compile/unmanagedClasspath/_global/streams/out
./target/streams/compile/compile
./target/streams/compile/compile/_global
./target/streams/compile/compile/_global/streams
./target/streams/compile/compile/_global/streams/out
./target/streams/compile/dependencyClasspath
./target/streams/compile/dependencyClasspath/_global
./target/streams/compile/dependencyClasspath/_global/streams
./target/streams/compile/dependencyClasspath/_global/streams/export
./target/streams/compile/unmanagedJars
./target/streams/compile/unmanagedJars/_global
./target/streams/compile/unmanagedJars/_global/streams
./target/streams/compile/unmanagedJars/_global/streams/export
./target/global-logging
./target/scala-2.12
./target/scala-2.12/sync
./target/scala-2.12/sync/copy-resource
./target/scala-2.12/zinc
./target/scala-2.12/zinc/inc_compile_2.12.zip
./target/scala-2.12/classes
./target/scala-2.12/classes/SocketStream$.class
./target/scala-2.12/classes/SocketStream.class
./target/scala-2.12/update
./target/scala-2.12/update/update_cache_2.12
./target/scala-2.12/update/update_cache_2.12/inputs
./target/scala-2.12/update/update_cache_2.12/output
./target/scala-2.12/socket-stream_2.12-1.0.jar
./target/task-temp-directory
./build.sbt
./src
./src/main
./src/main/scala
./src/main/scala/SocketStream.scala
```

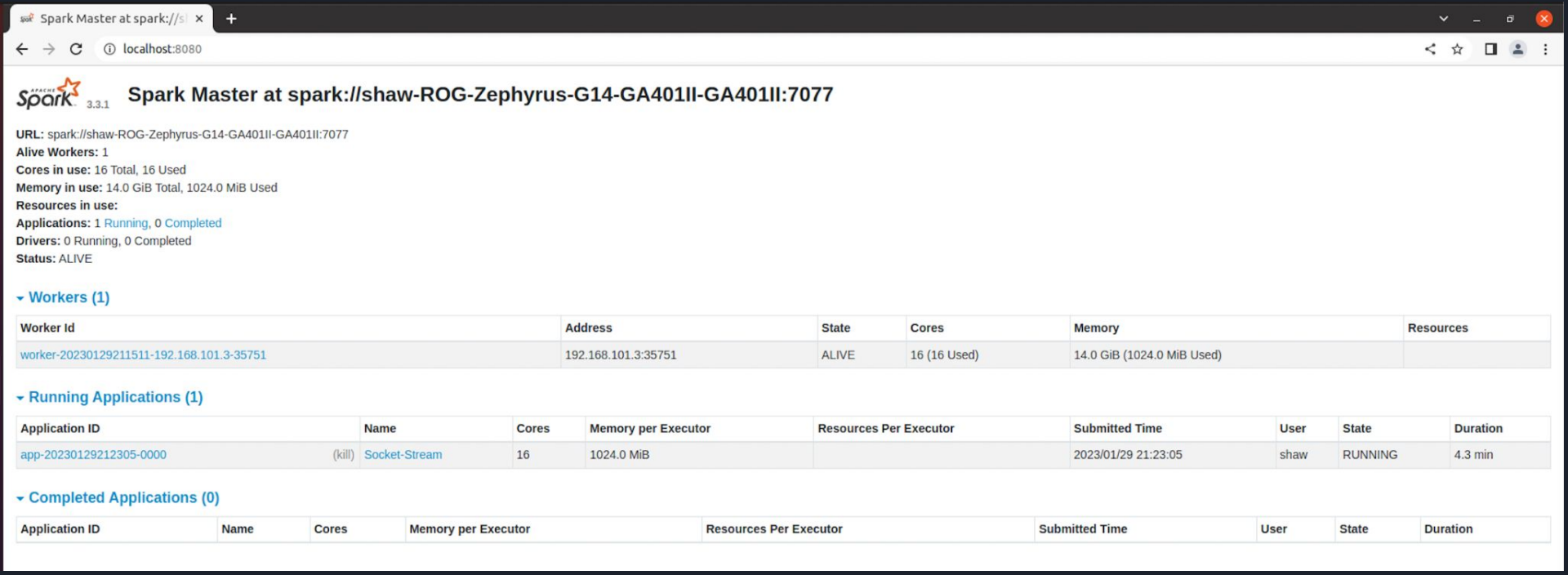
2. Socket Stream

Submit jar file to Master node

```
shaw@shaw-ROG-Zephyrus-G14-GA401II-GA401II:~/spark-practice/spark-3.3.1-bin-hadoop3/examples/socket-stream$ $SPARK_HOME/bin/spark-submit --master spark://shaw-ROG-Zephyrus-G14-GA401II-GA401II:7077 --class SocketStream target/scala-2.12/socket-stream_2.12-1.0.jar
23/01/29 21:23:04 WARN Utils: Your hostname, shaw-ROG-Zephyrus-G14-GA401II-GA401II resolves to a loopback address: 127.0.1.1; using 192.168.101.3 instead (on interface wlp2s0)
23/01/29 21:23:04 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
23/01/29 21:23:04 INFO SparkContext: Running Spark version 3.3.1
23/01/29 21:23:04 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
23/01/29 21:23:04 INFO ResourceUtils: ==============================================================
23/01/29 21:23:04 INFO ResourceUtils: No custom resources configured for spark.driver.
23/01/29 21:23:04 INFO ResourceUtils: ==============================================================
23/01/29 21:23:04 INFO SparkContext: Submitted application: Socket-Stream
23/01/29 21:23:04 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)
23/01/29 21:23:04 INFO ResourceProfile: Limiting resource is cpu
23/01/29 21:23:04 INFO ResourceProfileManager: Added ResourceProfile id: 0
23/01/29 21:23:04 INFO SecurityManager: Changing view acls to: shaw
23/01/29 21:23:04 INFO SecurityManager: Changing modify acls to: shaw
23/01/29 21:23:04 INFO SecurityManager: Changing view acls groups to:
23/01/29 21:23:04 INFO SecurityManager: Changing modify acls groups to:
23/01/29 21:23:04 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(shaw); groups with view permissions: Set(); users  with modify permissions: Set(shaw); groups with modify permissions: Set()
23/01/29 21:23:04 INFO Utils: Successfully started service 'sparkDriver' on port 40035.
23/01/29 21:23:04 INFO SparkEnv: Registering MapOutputTracker
23/01/29 21:23:04 INFO SparkEnv: Registering BlockManagerMaster
23/01/29 21:23:04 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
23/01/29 21:23:04 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
23/01/29 21:23:04 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
23/01/29 21:23:04 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-83038423-fb51-43c2-b18e-3c96f6f4cc8b
23/01/29 21:23:04 INFO MemoryStore: MemoryStore started with capacity 434.4 MiB
23/01/29 21:23:04 INFO SparkEnv: Registering OutputCommitCoordinator
23/01/29 21:23:05 INFO Utils: Successfully started service 'SparkUI' on port 4040.
```

2. Socket Stream

View Running Applications on UI: Socket-Stream



2. Socket Stream

Error connecting to port 7777 at the beginning

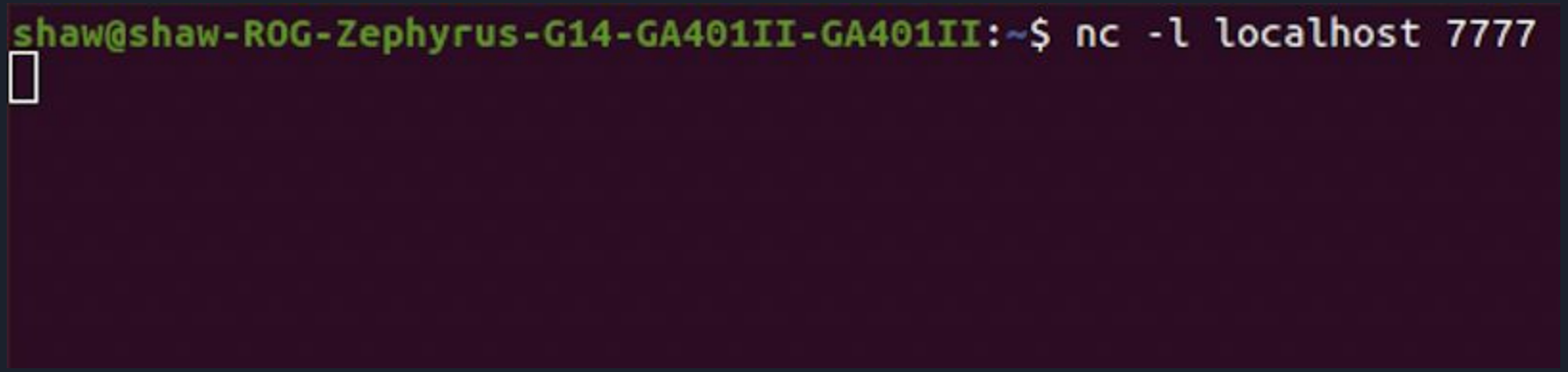
```
-------------------------------------------
Time: 1675002191000 ms
-------------------------------------------

23/01/29 21:23:11 INFO JobScheduler: Finished job streaming job 1675002191000 ms.0 from job set of time 1675002191000 ms
23/01/29 21:23:11 INFO JobScheduler: Total delay: 0.007 s for time 1675002191000 ms (execution: 0.001 s)
23/01/29 21:23:11 INFO MapPartitionsRDD: Removing RDD 5 from persistence list
23/01/29 21:23:11 INFO BlockManager: Removing RDD 5
23/01/29 21:23:11 INFO BlockRDD: Removing RDD 4 from persistence list
23/01/29 21:23:11 INFO BlockManager: Removing RDD 4
23/01/29 21:23:11 INFO SocketInputDStream: Removing blocks of RDD BlockRDD[4] at socketTextStream at SocketStream.scala:17 of time 1675002191000 ms
23/01/29 21:23:11 INFO ReceivedBlockTracker: Deleting batches:
23/01/29 21:23:11 INFO InputInfoTracker: remove old batch metadata:
23/01/29 21:23:11 INFO ReceiverTracker: Registered receiver for stream 0 from 192.168.101.3:39744
23/01/29 21:23:11 ERROR ReceiverTracker: Deregistered receiver for stream 0: Restarting receiver with delay 2000ms: Error connecting to localhost:7777 - java.net.ConnectException: Connection refused (Connection refused)
        at java.base/java.net.PlainSocketImpl.socketConnect(Native Method)
        at java.base/java.net.AbstractPlainSocketImpl.doConnect(AbstractPlainSocketImpl.java:412)
        at java.base/java.net.AbstractPlainSocketImpl.connectToAddress(AbstractPlainSocketImpl.java:255)
        at java.base/java.net.AbstractPlainSocketImpl.connect(AbstractPlainSocketImpl.java:237)
        at java.base/java.net.SocksSocketImpl.connect(SocksSocketImpl.java:392)
        at java.base/java.net.Socket.connect(Socket.java:609)
        at java.base/java.net.Socket.connect(Socket.java:558)
        at java.base/java.net.Socket.<init>(Socket.java:454)
        at java.base/java.net.Socket.<init>(Socket.java:231)
        at org.apache.spark.streaming.dstream.SocketReceiver.onStart(SocketInputDStream.scala:61)
        at org.apache.spark.streaming.receiver.ReceiverSupervisor.startReceiver(ReceiverSupervisor.scala:149)
        at org.apache.spark.streaming.receiver.ReceiverSupervisor.$anonfun$restartReceiver$1(ReceiverSupervisor.scala:198)
        at scala.runtime.java8.JFunction0$mcV$sp.apply(JFunction0$mcV$sp.java:23)
        at scala.concurrent.Future$.$anonfun$apply$1(Future.scala:659)
        at scala.util.Success.$anonfun$map$1(Try.scala:255)
        at scala.util.Success.map(Try.scala:213)
        at scala.concurrent.Future.$anonfun$map$1(Future.scala:292)
        at scala.concurrent.impl.Promise.liftedTree1$1(Promise.scala:33)
        at scala.concurrent.impl.Promise.$anonfun$transform$1(Promise.scala:33)
        at scala.concurrent.impl.CallbackRunnable.run(Promise.scala:64)
        at java.base/java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1128)
        at java.base/java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:628)
        at java.base/java.lang.Thread.run(Thread.java:829)

23/01/29 21:23:12 INFO JobScheduler: Added jobs for time 1675002192000 ms
23/01/29 21:23:12 INFO JobScheduler: Starting job streaming job 1675002192000 ms.0 from job set of time 1675002192000 ms
-------------------------------------------
Time: 1675002192000 ms
-------------------------------------------
```

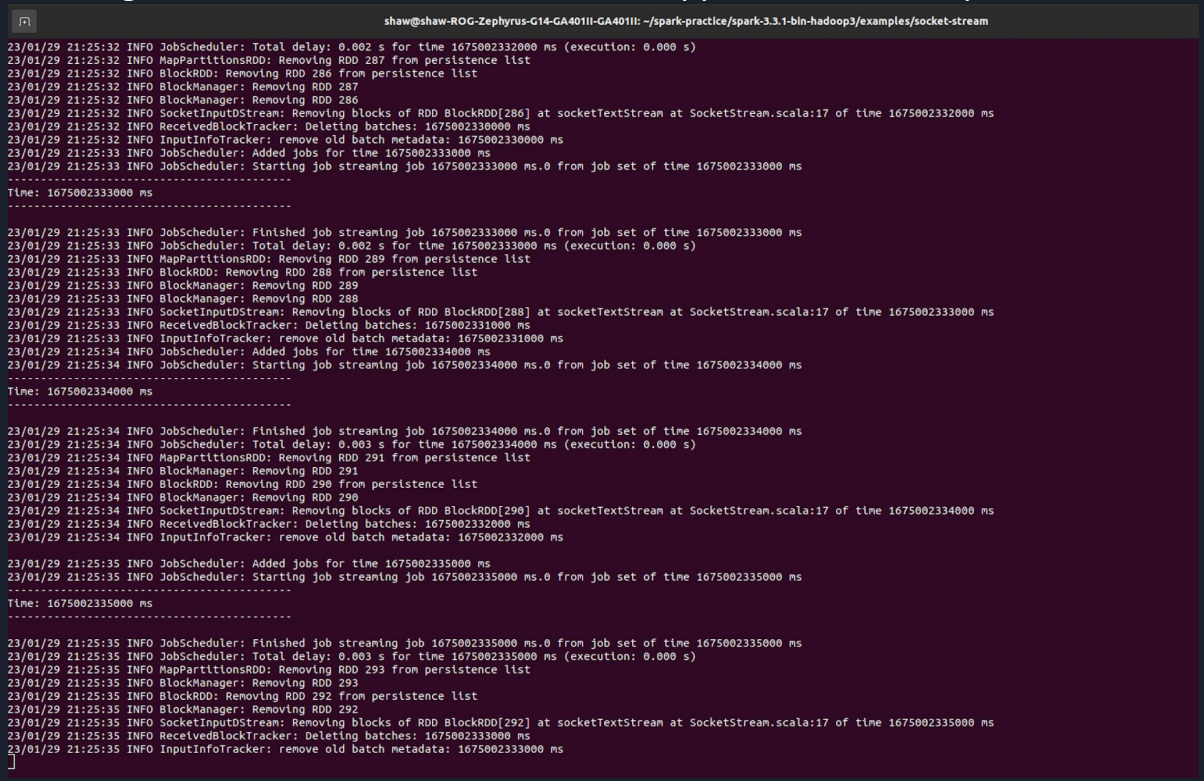
2. Socket Stream

Using netcat to communicate with port 7777



2. Socket Stream

No longer have connection error between application and the port 7777



2. Socket Stream - Test Case

Write to port 7777 first line contains the word “error”

```
shaw@shaw-ROG-Zephyrus-G14-GA401II-GA401II:~$ nc -l localhost 7777
{"status_code": 404, "status_text": "error", "message": "url not found"}
```

2. Socket Stream - Test Case

1 new job created, found the word “error”, the application print the whole line on the screen

```
-------------------------------------------
Time: 1675002721000 ms
-------------------------------------------

23/01/29 21:32:01 INFO JobScheduler: Finished job streaming job 1675002721000 ms.0 from job set of time 1675002721000 ms
23/01/29 21:32:01 INFO JobScheduler: Total delay: 0.002 s for time 1675002721000 ms (execution: 0.000 s)
23/01/29 21:32:01 INFO MapPartitionsRDD: Removing RDD 1065 from persistence list
23/01/29 21:32:01 INFO BlockRDD: Removing RDD 1064 from persistence list
23/01/29 21:32:01 INFO BlockManager: Removing RDD 1065
23/01/29 21:32:01 INFO SocketInputDStream: Removing blocks of RDD BlockRDD[1064] at socketTextStream at SocketStream.scala:17 of time 1675002721000 ms
23/01/29 21:32:01 INFO ReceivedBlockTracker: Deleting batches: 1675002719000 ms
23/01/29 21:32:01 INFO InputInfoTracker: remove old batch metadata: 1675002719000 ms
23/01/29 21:32:01 INFO BlockManager: Removing RDD 1064
23/01/29 21:32:01 INFO BlockManagerInfo: Added input-0-1675002721000 in memory on 192.168.101.3:41749 (size: 79.0 B, free: 434.4 MiB)
23/01/29 21:32:02 INFO JobScheduler: Added jobs for time 1675002722000 ms
23/01/29 21:32:02 INFO JobScheduler: Starting job streaming job 1675002722000 ms.0 from job set of time 1675002722000 ms
23/01/29 21:32:02 INFO SparkContext: Starting job: print at SocketStream.scala:23
23/01/29 21:32:02 INFO DAGScheduler: Got job 2 (print at SocketStream.scala:23) with 1 output partitions
23/01/29 21:32:02 INFO DAGScheduler: Final stage: ResultStage 3 (print at SocketStream.scala:23)
23/01/29 21:32:02 INFO DAGScheduler: Parents of final stage: List()
23/01/29 21:32:02 INFO DAGScheduler: Missing parents: List()
23/01/29 21:32:02 INFO DAGScheduler: Submitting ResultStage 3 (MapPartitionsRDD[1069] at filter at SocketStream.scala:20), which has no missing parents
23/01/29 21:32:02 INFO MemoryStore: Block broadcast_3 stored as values in memory (estimated size 3.7 KiB, free 434.3 MiB)
23/01/29 21:32:02 INFO MemoryStore: Block broadcast_3_piece0 stored as bytes in memory (estimated size 2.1 KiB, free 434.3 MiB)
23/01/29 21:32:02 INFO BlockManagerInfo: Added broadcast_3_piece0 in memory on 192.168.101.3:36715 (size: 2.1 KiB, free: 434.4 MiB)
23/01/29 21:32:02 INFO SparkContext: Created broadcast 3 from broadcast at DAGScheduler.scala:1513
23/01/29 21:32:02 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 3 (MapPartitionsRDD[1069] at filter at SocketStream.scala:20) (first 15 tasks are for partitions Vector(0))
23/01/29 21:32:02 INFO TaskSchedulerImpl: Adding task set 3.0 with 1 tasks resource profile 0
23/01/29 21:32:02 INFO TaskSetManager: Starting task 0.0 in stage 3.0 (TID 71) (192.168.101.3, executor 0, partition 0, PROCESS_LOCAL, 4398 bytes) taskResourceAssignments Map()
23/01/29 21:32:02 INFO BlockManagerInfo: Added broadcast_3_piece0 in memory on 192.168.101.3:41749 (size: 2.1 KiB, free: 434.4 MiB)
23/01/29 21:32:02 INFO TaskSetManager: Finished task 0.0 in stage 3.0 (TID 71) in 33 ms on 192.168.101.3 (executor 0) (1/1)
23/01/29 21:32:02 INFO TaskSchedulerImpl: Removed TaskSet 3.0, whose tasks have all completed, from pool
23/01/29 21:32:02 INFO DAGScheduler: ResultStage 3 (print at SocketStream.scala:23) finished in 0.040 s
23/01/29 21:32:02 INFO DAGScheduler: Job 2 is finished. Cancelling potential speculative or zombie tasks for this job
23/01/29 21:32:02 INFO TaskSchedulerImpl: Killing all running tasks in stage 3: Stage finished
23/01/29 21:32:02 INFO DAGScheduler: Job 2 finished: print at SocketStream.scala:23, took 0.045227 s
-------------------------------------------
Time: 1675002722000 ms
-------------------------------------------
{"status_code": 404, "status_text": "error", "message": "url not found"}
```

2. Socket Stream - Test Case

Write to port 7777 second line not contain the word “error”

```
shaw@shaw-ROG-Zephyrus-G14-GA401II-GA401II:~$ nc -l localhost 7777
{"status_code": 404, "status_text": "error", "message": "url not found"}
{"status_code": 200, "status_text": "success", "message": "OK"}
```

2. Socket Stream - Test Case

1 new job created, nothing printed on the screen

```
-------------------------------------------
Time: 1675002812000 ms
-------------------------------------------

23/01/29 21:33:32 INFO JobScheduler: Finished job streaming job 1675002812000 ms.0 from job set of time 1675002812000 ms
23/01/29 21:33:32 INFO JobScheduler: Total delay: 0.003 s for time 1675002812000 ms (execution: 0.001 s)
23/01/29 21:33:32 INFO MapPartitionsRDD: Removing RDD 1247 from persistence list
23/01/29 21:33:32 INFO BlockManager: Removing RDD 1247
23/01/29 21:33:32 INFO BlockRDD: Removing RDD 1246 from persistence list
23/01/29 21:33:32 INFO BlockManager: Removing RDD 1246
23/01/29 21:33:32 INFO SocketInputDStream: Removing blocks of RDD BlockRDD[1246] at socketTextStream at SocketStream.scala:17 of time 1675002812000 ms
23/01/29 21:33:32 INFO ReceivedBlockTracker: Deleting batches: 1675002810000 ms
23/01/29 21:33:32 INFO InputInfoTracker: remove old batch metadata: 1675002810000 ms
23/01/29 21:33:32 INFO BlockManagerInfo: Added input-0-1675002811800 in memory on 192.168.101.3:41749 (size: 70.0 B, free: 434.4 MiB)
23/01/29 21:33:33 INFO JobScheduler: Added jobs for time 1675002813000 ms
23/01/29 21:33:33 INFO JobScheduler: Starting job streaming job 1675002813000 ms.0 from job set of time 1675002813000 ms
23/01/29 21:33:33 INFO SparkContext: Starting job: print at SocketStream.scala:23
23/01/29 21:33:33 INFO DAGScheduler: Got job 3 (print at SocketStream.scala:23) with 1 output partitions
23/01/29 21:33:33 INFO DAGScheduler: Final stage: ResultStage 4 (print at SocketStream.scala:23)
23/01/29 21:33:33 INFO DAGScheduler: Parents of final stage: List()
23/01/29 21:33:33 INFO DAGScheduler: Missing parents: List()
23/01/29 21:33:33 INFO DAGScheduler: Submitting ResultStage 4 (MapPartitionsRDD[1251] at filter at SocketStream.scala:20), which has no missing parents
23/01/29 21:33:33 INFO MemoryStore: Block broadcast_4 stored as values in memory (estimated size 3.7 KiB, free 434.3 MiB)
23/01/29 21:33:33 INFO MemoryStore: Block broadcast_4_piece0 stored as bytes in memory (estimated size 2.1 KiB, free 434.3 MiB)
23/01/29 21:33:33 INFO BlockManagerInfo: Added broadcast_4_piece0 in memory on 192.168.101.3:36715 (size: 2.1 KiB, free: 434.4 MiB)
23/01/29 21:33:33 INFO SparkContext: Created broadcast 4 from broadcast at DAGScheduler.scala:1513
23/01/29 21:33:33 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 4 (MapPartitionsRDD[1251] at filter at SocketStream.scala:20) (first 15 tasks are for partitions Vector(0))
23/01/29 21:33:33 INFO TaskSchedulerImpl: Adding task set 4.0 with 1 tasks resource profile 0
23/01/29 21:33:33 INFO TaskSetManager: Starting task 0.0 in stage 4.0 (TID 72) (192.168.101.3, executor 0, partition 0, PROCESS_LOCAL, 4398 bytes) taskResourceAssignments Map()
23/01/29 21:33:33 INFO BlockManagerInfo: Added broadcast_4_piece0 in memory on 192.168.101.3:41749 (size: 2.1 KiB, free: 434.4 MiB)
23/01/29 21:33:33 INFO TaskSetManager: Finished task 0.0 in stage 4.0 (TID 72) in 15 ms on 192.168.101.3 (executor 0) (1/1)
23/01/29 21:33:33 INFO TaskSchedulerImpl: Removed TaskSet 4.0, whose tasks have all completed, from pool
23/01/29 21:33:33 INFO DAGScheduler: ResultStage 4 (print at SocketStream.scala:23) finished in 0.023 s
23/01/29 21:33:33 INFO DAGScheduler: Job 3 is finished. Cancelling potential speculative or zombie tasks for this job
23/01/29 21:33:33 INFO TaskSchedulerImpl: Killing all running tasks in stage 4: Stage finished
23/01/29 21:33:33 INFO DAGScheduler: Job 3 finished: print at SocketStream.scala:23, took 0.025906 s
-------------------------------------------
Time: 1675002813000 ms
-------------------------------------------
```

2. Socket Stream - Test Case

Write to port 7777 third line contains the word “Error” to check if the application check the uppercase letter “E”

```
shaw@shaw-ROG-Zephyrus-G14-GA401II-GA401II:~$ nc -l localhost 7777
{"status_code": 404, "status_text": "error", "message": "url not found"}
{"status_code": 200, "status_text": "success", "message": "OK"}
{"status_code": 404, "status_text": "Error", "message": "Url Not Found"}
```

2. Socket Stream - Test Case

1 new job created, nothing printed on the screen

```
-------------------------------------------
Time: 1675002866000 ms
-------------------------------------------

23/01/29 21:34:26 INFO JobScheduler: Finished job streaming job 1675002866000 ms.0 from job set of time 1675002866000 ms
23/01/29 21:34:26 INFO MapPartitionsRDD: Removing RDD 1355 from persistence list
23/01/29 21:34:26 INFO JobScheduler: Total delay: 0.002 s for time 1675002866000 ms (execution: 0.000 s)
23/01/29 21:34:26 INFO BlockManager: Removing RDD 1355
23/01/29 21:34:26 INFO BlockRDD: Removing RDD 1354 from persistence list
23/01/29 21:34:26 INFO BlockManager: Removing RDD 1354
23/01/29 21:34:26 INFO SocketInputDStream: Removing blocks of RDD BlockRDD[1354] at socketTextStream at SocketStream.scala:17 of time 1675002866000 ms
23/01/29 21:34:26 INFO ReceivedBlockTracker: Deleting batches: 1675002864000 ms
23/01/29 21:34:26 INFO InputInfoTracker: remove old batch metadata: 1675002864000 ms
23/01/29 21:34:26 INFO BlockManagerInfo: Added input-0-1675002866000 in memory on 192.168.101.3:41749 (size: 79.0 B, free: 434.4 MiB)
23/01/29 21:34:27 INFO JobScheduler: Added jobs for time 1675002867000 ms
23/01/29 21:34:27 INFO JobScheduler: Starting job streaming job 1675002867000 ms.0 from job set of time 1675002867000 ms
23/01/29 21:34:27 INFO SparkContext: Starting job: print at SocketStream.scala:23
23/01/29 21:34:27 INFO DAGScheduler: Got job 4 (print at SocketStream.scala:23) with 1 output partitions
23/01/29 21:34:27 INFO DAGScheduler: Final stage: ResultStage 5 (print at SocketStream.scala:23)
23/01/29 21:34:27 INFO DAGScheduler: Parents of final stage: List()
23/01/29 21:34:27 INFO DAGScheduler: Missing parents: List()
23/01/29 21:34:27 INFO DAGScheduler: Submitting ResultStage 5 (MapPartitionsRDD[1359] at filter at SocketStream.scala:20), which has no missing parents
23/01/29 21:34:27 INFO MemoryStore: Block broadcast_5 stored as values in memory (estimated size 3.7 KiB, free 434.3 MiB)
23/01/29 21:34:27 INFO MemoryStore: Block broadcast_5_piece0 stored as bytes in memory (estimated size 2.1 KiB, free 434.3 MiB)
23/01/29 21:34:27 INFO BlockManagerInfo: Added broadcast_5_piece0 in memory on 192.168.101.3:36715 (size: 2.1 KiB, free: 434.4 MiB)
23/01/29 21:34:27 INFO SparkContext: Created broadcast 5 from broadcast at DAGScheduler.scala:1513
23/01/29 21:34:27 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 5 (MapPartitionsRDD[1359] at filter at SocketStream.scala:20) (first 15 tasks are for partitions Vector(0))
23/01/29 21:34:27 INFO TaskSchedulerImpl: Adding task set 5.0 with 1 tasks resource profile 0
23/01/29 21:34:27 INFO TaskSetManager: Starting task 0.0 in stage 5.0 (TID 73) (192.168.101.3, executor 0, partition 0, PROCESS_LOCAL, 4398 bytes) taskResourceAssignments Map()
23/01/29 21:34:27 INFO BlockManagerInfo: Added broadcast_5_piece0 in memory on 192.168.101.3:41749 (size: 2.1 KiB, free: 434.4 MiB)
23/01/29 21:34:27 INFO TaskSetManager: Finished task 0.0 in stage 5.0 (TID 73) in 16 ms on 192.168.101.3 (executor 0) (1/1)
23/01/29 21:34:27 INFO TaskSchedulerImpl: Removed TaskSet 5.0, whose tasks have all completed, from pool
23/01/29 21:34:27 INFO DAGScheduler: ResultStage 5 (print at SocketStream.scala:23) finished in 0.023 s
23/01/29 21:34:27 INFO DAGScheduler: Job 4 is finished. Cancelling potential speculative or zombie tasks for this job
23/01/29 21:34:27 INFO TaskSchedulerImpl: Killing all running tasks in stage 5: Stage finished
23/01/29 21:34:27 INFO DAGScheduler: Job 4 finished: print at SocketStream.scala:23, took 0.026172 s
-------------------------------------------
Time: 1675002867000 ms
-------------------------------------------
```

2. Socket Stream

The Application detail UI

Spark Jobs (?)

User: shaw
Total Uptime: 13 min
Scheduling Mode: FIFO
Active Jobs: 1
Completed Jobs: 4

▸ Event Timeline

▾ Active Jobs (1)

Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page. Go

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
1	Streaming job running receiver 0 start at SocketStream.scala:26 (kill)	2023/01/29 21:23:09	13 min	0/1	0/1 (1 running)

Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page. Go

▾ Completed Jobs (4)

Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page. Go

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
4	Streaming job from [output operation 0, batch time 21:34:27] print at SocketStream.scala:23	2023/01/29 21:34:27	25 ms	1/1	1/1
3	Streaming job from [output operation 0, batch time 21:33:33] print at SocketStream.scala:23	2023/01/29 21:33:33	25 ms	1/1	1/1
2	Streaming job from [output operation 0, batch time 21:32:02] print at SocketStream.scala:23	2023/01/29 21:32:02	44 ms	1/1	1/1
0	start at SocketStream.scala:26 start at SocketStream.scala:26	2023/01/29 21:23:06	3 s	2/2	70/70

Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page. Go

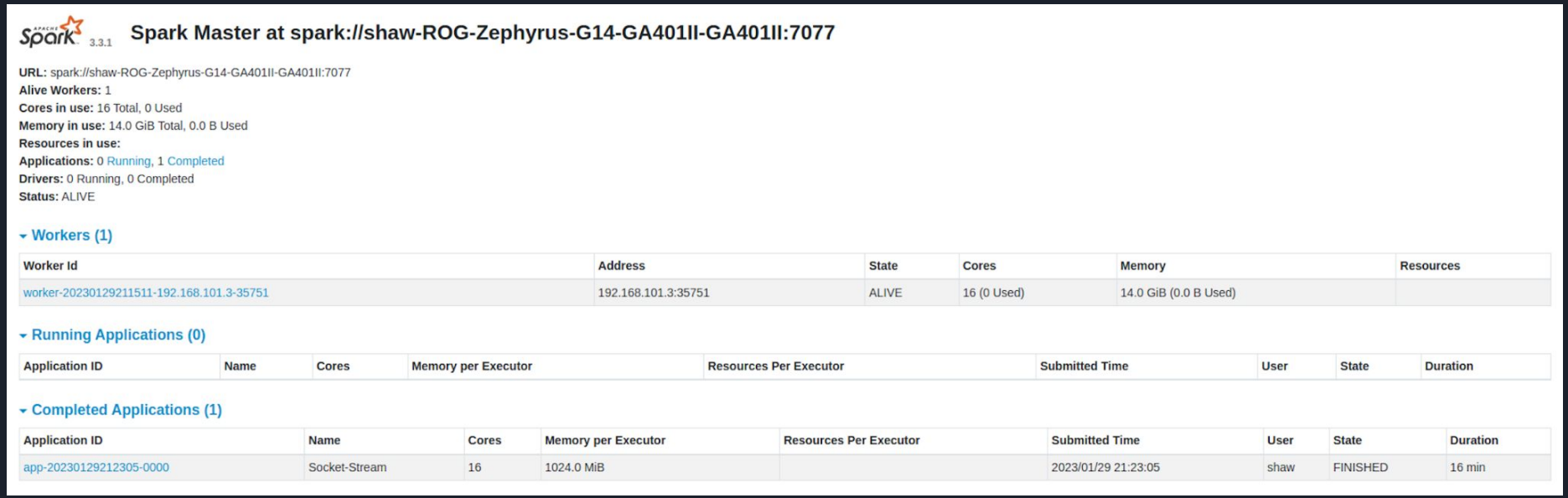
2. Socket Stream

Stop the application

```
23/01/29 21:39:04 INFO DAGScheduler: Executor lost: 0 (epoch 1)
23/01/29 21:39:04 INFO BlockManagerMasterEndpoint: Trying to remove executor 0 from BlockManagerMaster.
23/01/29 21:39:04 INFO BlockManagerMasterEndpoint: Removing block manager BlockManagerId(0, 192.168.101.3, 41749, None)
23/01/29 21:39:04 INFO SparkUI: Stopped Spark web UI at http://192.168.101.3:4040
23/01/29 21:39:04 INFO BlockManagerMaster: Removed 0 successfully in removeExecutor
23/01/29 21:39:04 INFO DAGScheduler: Shuffle files lost for executor: 0 (epoch 1)
23/01/29 21:39:04 INFO StandaloneSchedulerBackend: Shutting down all executors
23/01/29 21:39:04 INFO CoarseGrainedSchedulerBackend$DriverEndpoint: Asking each executor to shut down
23/01/29 21:39:04 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
23/01/29 21:39:04 INFO MemoryStore: MemoryStore cleared
23/01/29 21:39:04 INFO BlockManager: BlockManager stopped
23/01/29 21:39:04 INFO BlockManagerMaster: BlockManagerMaster stopped
23/01/29 21:39:04 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
23/01/29 21:39:04 INFO SparkContext: Successfully stopped SparkContext
23/01/29 21:39:04 INFO ShutdownHookManager: Shutdown hook called
23/01/29 21:39:04 INFO ShutdownHookManager: Deleting directory /tmp/spark-236a16fa-3e19-4ce3-9da9-9fa207d54d37
23/01/29 21:39:04 INFO ShutdownHookManager: Deleting directory /tmp/spark-46f8e96f-e877-4b13-9f86-7a781c0f78a6
shaw@shaw-ROG-Zephyrus-G14-GA401II-GA401II:~/spark-practice/spark-3.3.1-bin-hadoop3/examples/socket-stream$
```

2. Socket Stream

View the UI, the application was completed



Spark 3.3.1

Spark Master at spark://shaw-ROG-Zephyrus-G14-GA401II-GA401II:7077

URL: spark://shaw-ROG-Zephyrus-G14-GA401II-GA401II:7077
Alive Workers: 1
Cores in use: 16 Total, 0 Used
Memory in use: 14.0 GiB Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 1 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (1)

Worker Id	Address	State	Cores	Memory	Resources
worker-20230129211511-192.168.101.3-35751	192.168.101.3:35751	ALIVE	16 (0 Used)	14.0 GiB (0.0 B Used)	

Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

Completed Applications (1)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20230129212305-0000	Socket-Stream	16	1024.0 MiB		2023/01/29 21:23:05	shaw	FINISHED	16 min

3. Log Analyzer

Source Code: LogAnalyzerStreaming.scala

LogAnalyzerStreaming.scala

```scala
import org.apache.spark.SparkConf
import org.apache.spark.rdd.RDD
import org.apache.spark.streaming.dstream.DStream
import org.apache.spark.streaming.{ Seconds, StreamingContext }

object LogAnalyzerStreaming {
  def main(args: Array[String]) {
    val WINDOW_LENGTH = Seconds(30)
    val SLIDE_INTERVAL = Seconds(10)

    val sparkConf = new SparkConf().setAppName("Log Analyzer Streaming in Scala")
    val streamingContext = new StreamingContext(sparkConf, SLIDE_INTERVAL)
    val logLinesDStream: DStream[String] = streamingContext.socketTextStream("localhost", 9999)
    val accessLogsDStream: DStream[ApacheAccessLog] = logLinesDStream.map(ApacheAccessLog_Object.parseLogLine).cache()
    val windowDStream: DStream[ApacheAccessLog] = accessLogsDStream.window(WINDOW_LENGTH, SLIDE_INTERVAL)

    windowDStream.foreachRDD(accessLogs => {
      if (accessLogs.count() == 0) {
        println("No access logs received in this time interval")
      } else {
        // Calculate statistics based on the content size
        val contentSizes: RDD[Long] = accessLogs.map(_.contentSize).cache()
        println("Content Size Avg: %s, Min: %s, Max: %s".format(
          contentSizes.reduce(_ + _) / contentSizes.count,
          contentSizes.min,
          contentSizes.max))

        // Compute Response Code to Count
        val responseCodeToCount: Array[(Int, Long)] = accessLogs.map(_.responseCode -> 1L).reduceByKey(_ + _).take(100)
        println(s"""Response code counts: ${responseCodeToCount.mkString("[", ",", "]")}""")

        // Any IPAddress that has accessed the server more than 10 times
        val ipAddresses: Array[String] = accessLogs.map(_.ipAddress -> 1L).reduceByKey(_ + _).filter(_._2 > 10).map(_._1).take(100)
        println(s"""IPAddresses > 10 times: ${ipAddresses.mkString("[", ",", "]")}""")

        // Top Endpoints
        val topEndpoints: Array[(String, Long)] = accessLogs.map(_.endpoint -> 1L).reduceByKey(_ + _).top(10)(Ordering.by[(String, Long), Long](_._2))
        println(s"""Top Endpoints: ${topEndpoints.mkString("[", ",", "]")}""")
      }
    })

    //// Start the streaming server

    // Start the computation
    streamingContext.start()

    // Wait for the computation to terminate
    streamingContext.awaitTermination()
  }
}
```

3. Log Analyzer

Source Code:
ApacheAccessLog.scala

ApacheAccessLog.scala

```scala
/** An entry of Apache access log */

case class ApacheAccessLog(
  ipAddress: String,
  clientIdentd: String,
  userId: String,
  dateTime: String,
  method: String,
  endpoint: String,
  protocol: String,
  responseCode: Int,
  contentSize: Long) {
}

object ApacheAccessLog_Object {
  val PATTERN = """^(\S+) (\S+) (\S+) \[([\w:/]+\s[+\-]\d{4})\] "(\S+) (\S+) (\S+)" (\d{3}) (\d+)""".r
  /**
   * Parse log entry from a string.
   *
   * @param log A string, typically a line from a log file
   * @return An entry of Apache access log
   * @throws RuntimeException Unable to parse the string
   */

  def parseLogLine(log: String): ApacheAccessLog = {
    log match {
      case PATTERN(ipAddress, clientIdentd, userId, dateTime, method, endpoint, protocol, responseCode, contentSize) => ApacheAccessLog(ipAddress, clientIdentd,
      case _ => throw new RuntimeException(s"""Cannot parse log line: $log""")
    }
  }
}
```

3. Log Analyzer

Source Code: stream.sh

```
#!/bin/sh

set -o nounset
set -o errexit

test $# -eq 1 || (echo "Incorrect number of arguments"; exit 1)

file="$1"

network_port=9999
lines_in_batch=100
interval_sec=10

n_lines=$(cat $file | wc -l)
cursor=1

while test $cursor -le $n_lines
do
    tail -n +$cursor $file | head -$lines_in_batch | nc -l $network_port
    cursor=$(($cursor + $lines_in_batch))
    sleep $interval_sec
done

```

3. Log Analyzer

Setting: build.sbt

build.sbt

```
1  name := "logs-analyzer"
2
3  version := "1.0"
4
5  scalaVersion := "2.12.15"
6
7  libraryDependencies ++= Seq {
8    "org.apache.spark" %% "spark-core" % "3.3.1" % "provided";
9    "org.apache.spark" %% "spark-streaming" % "3.3.1"
10 }
11
```

3. Log Analyzer

Package the project by sbt

```
shaw@shaw-ROG-Zephyrus-G14-GA401II-GA401II:~/spark-practice/spark-3.3.1-bin-hadoop3/examples/logs-analyzer$ find .
.
./build.sbt
./src
./src/main
./src/main/scala
./src/main/scala/LogAnalyzerStreaming.scala
./src/main/scala/ApacheAccessLog.scala
shaw@shaw-ROG-Zephyrus-G14-GA401II-GA401II:~/spark-practice/spark-3.3.1-bin-hadoop3/examples/logs-analyzer$ sbt clean package
[info] Updated file /home/shaw/spark-practice/spark-3.3.1-bin-hadoop3/examples/logs-analyzer/project/build.properties: set sbt.version to 1.8.2
[info] welcome to sbt 1.8.2 (Ubuntu Java 11.0.17)
[info] loading project definition from /home/shaw/spark-practice/spark-3.3.1-bin-hadoop3/examples/logs-analyzer/project
[info] loading settings for project logs-analyzer from build.sbt ...
[info] set current project to logs-analyzer (in build file:/home/shaw/spark-practice/spark-3.3.1-bin-hadoop3/examples/logs-analyzer/)
[success] Total time: 0 s, completed Jan 29, 2023, 9:54:06 PM
[info] compiling 2 Scala sources to /home/shaw/spark-practice/spark-3.3.1-bin-hadoop3/examples/logs-analyzer/target/scala-2.12/classes ...
[success] Total time: 6 s, completed Jan 29, 2023, 9:54:12 PM
```

3. Log Analyzer

Obtain jar file

- A JAR ("Java archive") is a package file format typically used to aggregate many Java class files and associated metadata resources (text, images, etc.) into one file for distribution

```
./target/streams/compile/internalDependencyClasspath/_global/streams/export
./target/streams/compile/internalDependencyClasspath/_global/streams/out
./target/streams/compile/_global
./target/streams/compile/_global/_global
./target/streams/compile/_global/_global/discoveredMainClasses
./target/streams/compile/_global/_global/discoveredMainClasses/data
./target/streams/compile/_global/_global/compileOutputs
./target/streams/compile/_global/_global/compileOutputs/previous
./target/streams/compile/bspReporter
./target/streams/compile/bspReporter/_global
./target/streams/compile/bspReporter/_global/streams
./target/streams/compile/bspReporter/_global/streams/out
./target/streams/compile/unmanagedClasspath
./target/streams/compile/unmanagedClasspath/_global
./target/streams/compile/unmanagedClasspath/_global/streams
./target/streams/compile/unmanagedClasspath/_global/streams/export
./target/streams/compile/unmanagedClasspath/_global/streams/out
./target/streams/compile/compile
./target/streams/compile/compile/_global
./target/streams/compile/compile/_global/streams
./target/streams/compile/compile/_global/streams/out
./target/streams/compile/dependencyClasspath
./target/streams/compile/dependencyClasspath/_global
./target/streams/compile/dependencyClasspath/_global/streams
./target/streams/compile/dependencyClasspath/_global/streams/export
./target/streams/compile/unmanagedJars
./target/streams/compile/unmanagedJars/_global
./target/streams/compile/unmanagedJars/_global/streams
./target/streams/compile/unmanagedJars/_global/streams/export
./target/global-logging
./target/scala-2.12
./target/scala-2.12/logs-analyzer_2.12-1.0.jar
./target/scala-2.12/sync
./target/scala-2.12/sync/copy-resource
./target/scala-2.12/zinc
./target/scala-2.12/zinc/inc_compile_2.12.zip
./target/scala-2.12/classes
./target/scala-2.12/classes/ApacheAccessLog.class
./target/scala-2.12/classes/ApacheAccessLog_Object.class
./target/scala-2.12/classes/ApacheAccessLog$.class
./target/scala-2.12/classes/LogAnalyzerStreaming.class
./target/scala-2.12/classes/ApacheAccessLog_Object$.class
./target/scala-2.12/classes/LogAnalyzerStreaming$.class
./target/scala-2.12/update
./target/scala-2.12/update/update_cache_2.12
./target/scala-2.12/update/update_cache_2.12/inputs
./target/scala-2.12/update/update_cache_2.12/output
./target/task-temp-directory
./build.sbt
./src
./src/main
./src/main/scala
./src/main/scala/LogAnalyzerStreaming.scala
./src/main/scala/ApacheAccessLog.scala
shaw@shaw-ROG-Zephyrus-G14-GA401II-GA401II:~/spark-practice/spark-3.3.1-bin-hadoop3/examples/logs-analyzer$
```

3. Log Analyzer

Submit jar file to Master node

```
shaw@shaw-ROG-Zephyrus-G14-GA401II-GA401II:~/spark-practice/spark-3.3.1-bin-hadoop3/examples/logs-analyzer$ $SPARK_HOME/bin/spark-submit --master spark://shaw-ROG-Zephyrus-G14-GA401II-GA401II:7077 --class LogAnalyzerStreaming target/scala-2.12/logs-analyzer_2.12-1.0.jar
23/01/29 22:41:54 WARN Utils: Your hostname, shaw-ROG-Zephyrus-G14-GA401II-GA401II resolves to a loopback address: 127.0.1.1; using 192.168.101.3 instead (on interface wlp2s0)
23/01/29 22:41:54 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
23/01/29 22:42:03 INFO SparkContext: Running Spark version 3.3.1
23/01/29 22:42:03 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
23/01/29 22:42:03 INFO ResourceUtils: ==============================================================
23/01/29 22:42:03 INFO ResourceUtils: No custom resources configured for spark.driver.
23/01/29 22:42:03 INFO ResourceUtils: ==============================================================
23/01/29 22:42:03 INFO SparkContext: Submitted application: Log Analyzer Streaming in Scala
23/01/29 22:42:03 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)
23/01/29 22:42:03 INFO ResourceProfile: Limiting resource is cpu
23/01/29 22:42:03 INFO ResourceProfileManager: Added ResourceProfile id: 0
23/01/29 22:42:03 INFO SecurityManager: Changing view acls to: shaw
23/01/29 22:42:03 INFO SecurityManager: Changing modify acls to: shaw
23/01/29 22:42:03 INFO SecurityManager: Changing view acls groups to:
23/01/29 22:42:03 INFO SecurityManager: Changing modify acls groups to:
23/01/29 22:42:03 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(shaw); groups with view permissions: Set(); users  with modify permissions: Set(shaw); groups with modify permissions: Set()
23/01/29 22:42:04 INFO Utils: Successfully started service 'sparkDriver' on port 45397.
23/01/29 22:42:04 INFO SparkEnv: Registering MapOutputTracker
23/01/29 22:42:04 INFO SparkEnv: Registering BlockManagerMaster
23/01/29 22:42:04 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
23/01/29 22:42:04 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
23/01/29 22:42:04 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
23/01/29 22:42:04 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-d02ed0d1-77a0-481d-8778-78f67db7615c
23/01/29 22:42:04 INFO MemoryStore: MemoryStore started with capacity 434.4 MiB
23/01/29 22:42:04 INFO SparkEnv: Registering OutputCommitCoordinator
23/01/29 22:42:04 INFO Utils: Successfully started service 'SparkUI' on port 4040.
```

3. Log Analyzer

View Running Applications on UI: Log Analyzer Streaming in Scala

Spark 3.3.1

Spark Master at spark://shaw-ROG-Zephyrus-G14-GA401II-GA401II:7077

URL: spark://shaw-ROG-Zephyrus-G14-GA401II-GA401II:7077
Alive Workers: 1
Cores in use: 16 Total, 16 Used
Memory in use: 14.0 GiB Total, 1024.0 MiB Used
Resources in use:
Applications: 1 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (1)

Worker Id	Address	State	Cores	Memory	Resources
worker-20230129224017-192.168.101.3-43555	192.168.101.3:43555	ALIVE	16 (16 Used)	14.0 GiB (1024.0 MiB Used)	

Running Applications (1)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20230129224204-0000 (kill)	Log Analyzer Streaming in Scala	16	1024.0 MiB		2023/01/29 22:42:04	shaw	RUNNING	9 s

Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

3. Log Analyzer

Error connecting to port 9999 at the beginning

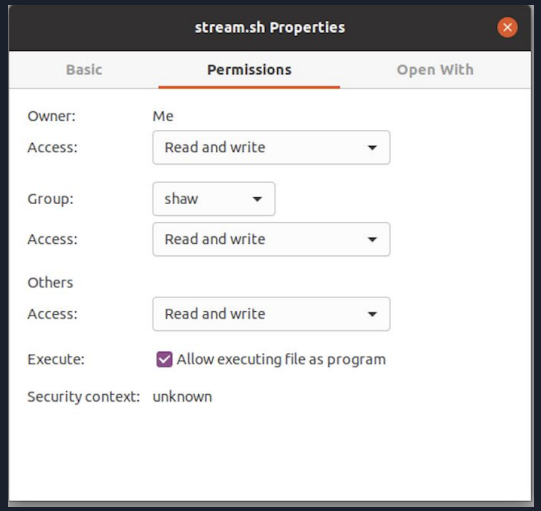
```
        at org.apache.spark.SparkContext.$anonfun$submitJob$1(SparkContext.scala:2377)
        at org.apache.spark.scheduler.ResultTask.runTask(ResultTask.scala:90)
        at org.apache.spark.scheduler.Task.run(Task.scala:136)
        at org.apache.spark.executor.Executor$TaskRunner.$anonfun$run$3(Executor.scala:548)
        at org.apache.spark.util.Utils$.tryWithSafeFinally(Utils.scala:1504)
        at org.apache.spark.executor.Executor$TaskRunner.run(Executor.scala:551)
        at java.base/java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1128)
        at java.base/java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:628)
        at java.base/java.lang.Thread.run(Thread.java:829)

23/01/29 22:42:18 INFO ReceiverTracker: Registered receiver for stream 0 from 192.168.101.3:57706
23/01/29 22:42:18 ERROR ReceiverTracker: Deregistered receiver for stream 0: Restarting receiver with delay 2000ms: Error connecting to localhost:9999 - java.net.ConnectException: Connection refused (Con
nection refused)
        at java.base/java.net.PlainSocketImpl.socketConnect(Native Method)
        at java.base/java.net.AbstractPlainSocketImpl.doConnect(AbstractPlainSocketImpl.java:412)
        at java.base/java.net.AbstractPlainSocketImpl.connectToAddress(AbstractPlainSocketImpl.java:255)
        at java.base/java.net.AbstractPlainSocketImpl.connect(AbstractPlainSocketImpl.java:237)
        at java.base/java.net.SocksSocketImpl.connect(SocksSocketImpl.java:392)
        at java.base/java.net.Socket.connect(Socket.java:609)
        at java.base/java.net.Socket.connect(Socket.java:558)
        at java.base/java.net.Socket.<init>(Socket.java:454)
        at java.base/java.net.Socket.<init>(Socket.java:231)
        at org.apache.spark.streaming.dstream.SocketReceiver.onStart(SocketInputDStream.scala:61)
        at org.apache.spark.streaming.receiver.ReceiverSupervisor.startReceiver(ReceiverSupervisor.scala:149)
        at org.apache.spark.streaming.receiver.ReceiverSupervisor.$anonfun$restartReceiver$1(ReceiverSupervisor.scala:198)
        at scala.runtime.java8.JFunction0$mcV$sp.apply(JFunction0$mcV$sp.java:23)
        at scala.concurrent.Future$.$anonfun$apply$1(Future.scala:659)
        at scala.util.Success.$anonfun$map$1(Try.scala:255)
        at scala.util.Success.map(Try.scala:213)
        at scala.concurrent.Future.$anonfun$map$1(Future.scala:292)
        at scala.concurrent.impl.Promise.liftedTree1$1(Promise.scala:33)
        at scala.concurrent.impl.Promise.$anonfun$transform$1(Promise.scala:33)
        at scala.concurrent.impl.CallbackRunnable.run(Promise.scala:64)
        at java.base/java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1128)
        at java.base/java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:628)
        at java.base/java.lang.Thread.run(Thread.java:829)
```

3. Log Analyzer

Run shell script: stream.sh to write logs from file logs.txt

Allow executing file stream.sh as program

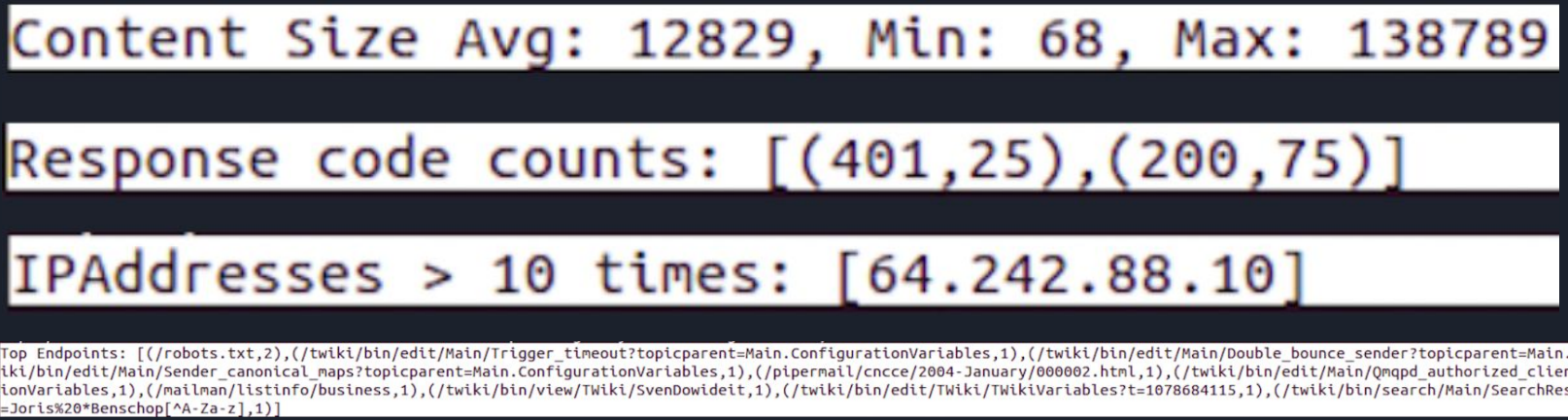


Run stream.sh

```
shaw@shaw-ROG-Zephyrus-G14-GA401II-GA401II:~/spark-practice/spark-3.3.1-bin-hadoop3/examples/logs-stream$ ./stream.sh logs.txt
```

3. Log Analyzer

While the shell was executing, the application read the logs, process and the results printed on the screen



3. Log Analyzer

When receive no logs

```
No access logs received in this time interval
```

was printed

3. Log Analyzer

The Jobs UI

	take at LogAnalyzerStreaming.scala:33				
17	Streaming job from [output operation 0, batch time 22:42:50] take at LogAnalyzerStreaming.scala:29	2023/01/29 22:42:50	64 ms	2/2	2/2
16	Streaming job from [output operation 0, batch time 22:42:50] max at LogAnalyzerStreaming.scala:26	2023/01/29 22:42:50	26 ms	1/1	1/1
15	Streaming job from [output operation 0, batch time 22:42:50] min at LogAnalyzerStreaming.scala:25	2023/01/29 22:42:50	28 ms	1/1	1/1
14	Streaming job from [output operation 0, batch time 22:42:50] count at LogAnalyzerStreaming.scala:24	2023/01/29 22:42:50	22 ms	1/1	1/1
13	Streaming job from [output operation 0, batch time 22:42:50] reduce at LogAnalyzerStreaming.scala:24	2023/01/29 22:42:50	34 ms	1/1	1/1
12	Streaming job from [output operation 0, batch time 22:42:50] count at LogAnalyzerStreaming.scala:18	2023/01/29 22:42:50	28 ms	1/1	1/1
11	Streaming job from [output operation 0, batch time 22:42:40] top at LogAnalyzerStreaming.scala:37	2023/01/29 22:42:40	0.2 s	2/2	2/2
10	Streaming job from [output operation 0, batch time 22:42:40] take at LogAnalyzerStreaming.scala:33	2023/01/29 22:42:40	87 ms	2/2	2/2
9	Streaming job from [output operation 0, batch time 22:42:40] take at LogAnalyzerStreaming.scala:29	2023/01/29 22:42:40	81 ms	2/2	2/2
8	Streaming job from [output operation 0, batch time 22:42:40] max at LogAnalyzerStreaming.scala:26	2023/01/29 22:42:40	33 ms	1/1	1/1
7	Streaming job from [output operation 0, batch time 22:42:40] min at LogAnalyzerStreaming.scala:25	2023/01/29 22:42:40	31 ms	1/1	1/1
6	Streaming job from [output operation 0, batch time 22:42:40] count at LogAnalyzerStreaming.scala:24	2023/01/29 22:42:40	37 ms	1/1	1/1
5	Streaming job from [output operation 0, batch time 22:42:40] reduce at LogAnalyzerStreaming.scala:24	2023/01/29 22:42:40	48 ms	1/1	1/1
4	Streaming job from [output operation 0, batch time 22:42:40] count at LogAnalyzerStreaming.scala:18	2023/01/29 22:42:40	73 ms	1/1	1/1
3	Streaming job from [output operation 0, batch time 22:42:30]	2023/01/29 22:42:30	0 ms	0/0	0/0
2	Streaming job from [output operation 0, batch time 22:42:20]	2023/01/29 22:42:20	0 ms	0/0	0/0
0	start at LogAnalyzerStreaming.scala:45 start at LogAnalyzerStreaming.scala:45	2023/01/29 22:42:05	11 s	2/2	70/70

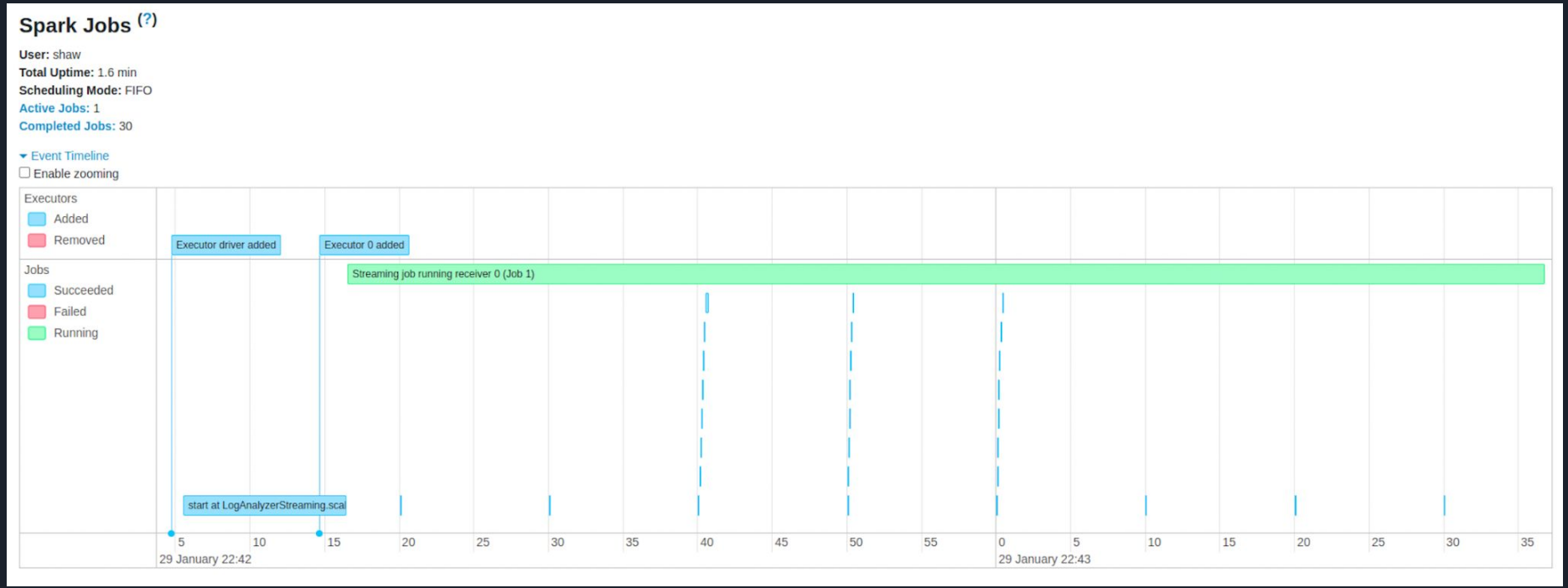
3. Log Analyzer

The Jobs UI

23	Streaming job from [output operation 0, batch time 22:43:00] min at LogAnalyzerStreaming.scala:25	2023/01/29 22:43:00	20 ms	1/1	1/1
22	Streaming job from [output operation 0, batch time 22:43:00] count at LogAnalyzerStreaming.scala:24	2023/01/29 22:43:00	21 ms	1/1	1/1
21	Streaming job from [output operation 0, batch time 22:43:00] reduce at LogAnalyzerStreaming.scala:24	2023/01/29 22:43:00	27 ms	1/1	1/1
20	Streaming job from [output operation 0, batch time 22:43:00] count at LogAnalyzerStreaming.scala:18	2023/01/29 22:43:00	24 ms	1/1	1/1
19	Streaming job from [output operation 0, batch time 22:42:50] top at LogAnalyzerStreaming.scala:37	2023/01/29 22:42:50	63 ms	2/2	2/2
18	Streaming job from [output operation 0, batch time 22:42:50] take at LogAnalyzerStreaming.scala:33	2023/01/29 22:42:50	57 ms	2/2	2/2
17	Streaming job from [output operation 0, batch time 22:42:50] take at LogAnalyzerStreaming.scala:29	2023/01/29 22:42:50	64 ms	2/2	2/2

3. Log Analyzer

The Event Timeline UI



This is end of our report slides!
Thank you for watching!