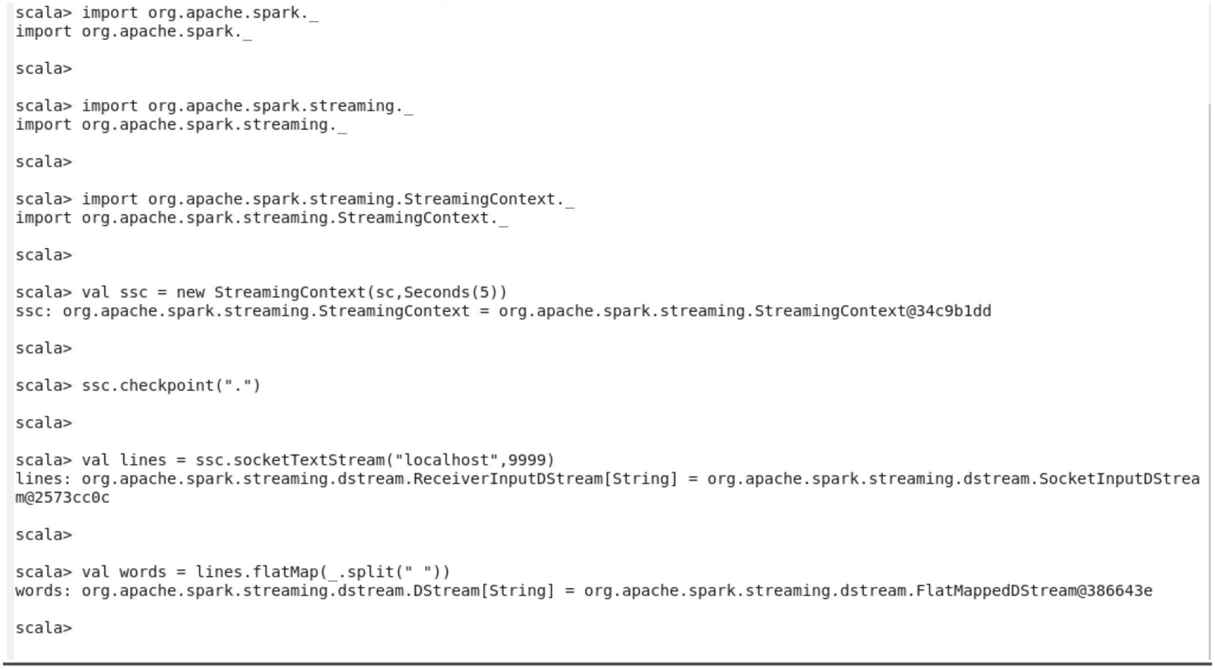
Problem Statement:

Implement the below blog at your end and send the complete documentation.
https://acadgild.com/blog/stateful-streaming-in-spark/

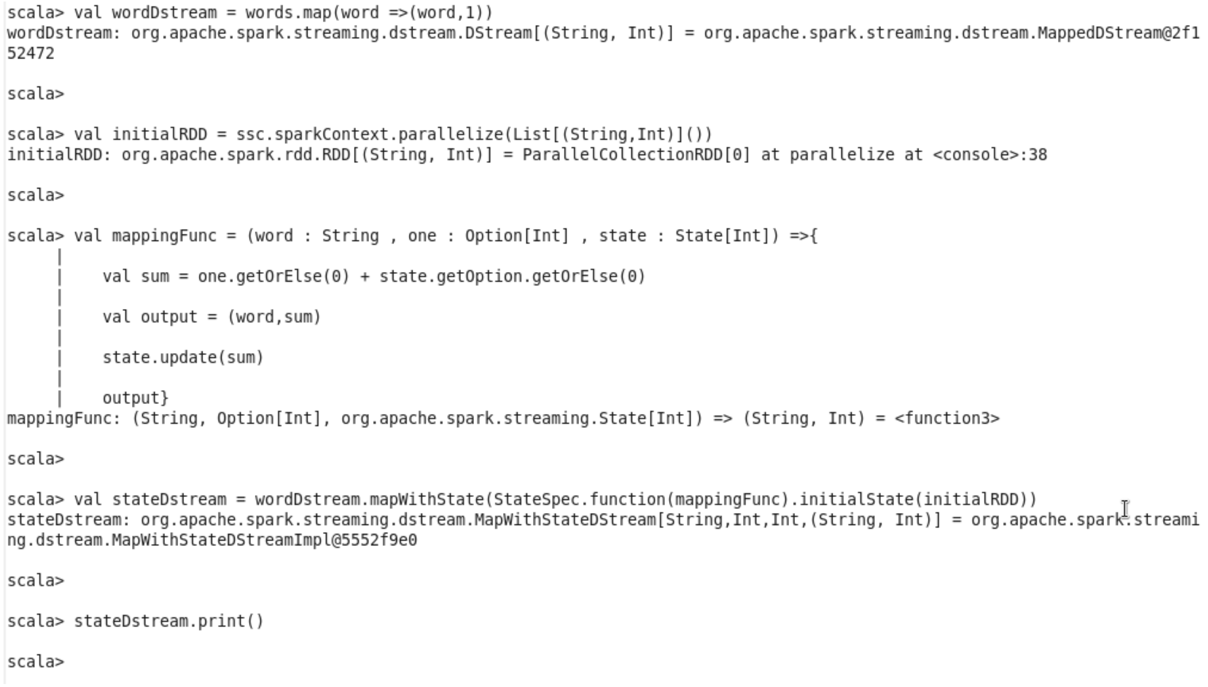
Solution:

- importing the spark packages
- Declaring the Streaming Context Object with the help of SPARK context object sc with batch interval 5 secs
- Setting the checkpoint to current home directory
- setting the host details for source input stream
- converting the data into words via flatmap operation
- Map operation for converting each word as key and 1 as value
- Parallelize operation for entire data received in a batch interval for the string in form of List
- Mapping function to get the count from system variable STATE and adding the same to current state of batch interval
- Applying the mapWithState function to current stream for stateful streaming
- Printing the data stream with word count
- Starting the SPARK STREAMING session
- Keep awaiting for the streamed data until streaming session is terminated with CTRL+C

Code:



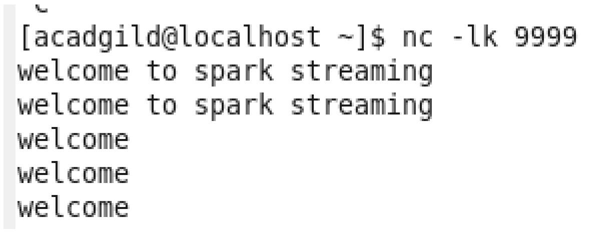
```
scala> import org.apache.spark._
import org.apache.spark._

scala>

scala> import org.apache.spark.streaming._
import org.apache.spark.streaming._

scala>

scala> import org.apache.spark.streaming.StreamingContext._
import org.apache.spark.streaming.StreamingContext._

scala>

scala> val ssc = new StreamingContext(sc,Seconds(5))
ssc: org.apache.spark.streaming.StreamingContext = org.apache.spark.streaming.StreamingContext@34c9b1dd

scala>

scala> ssc.checkpoint(".")

scala>

scala> val lines = ssc.socketTextStream("localhost",9999)
lines: org.apache.spark.streaming.dstream.ReceiverInputDStream[String] = org.apache.spark.streaming.dstream.SocketInputDStrea
m@2573cc0c

scala>

scala> val words = lines.flatMap(_.split(" "))
words: org.apache.spark.streaming.dstream.DStream[String] = org.apache.spark.streaming.dstream.FlatMappedDStream@386643e

scala>
```



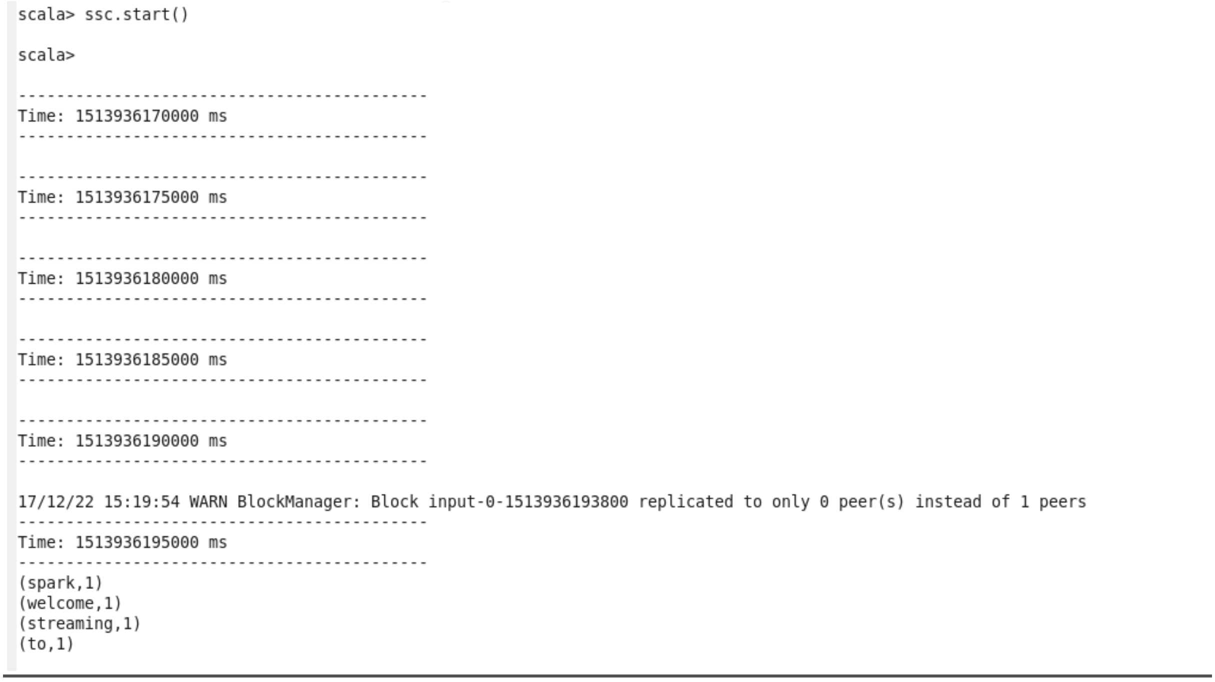
```
scala> val wordDstream = words.map(word =>(word,1))
wordDstream: org.apache.spark.streaming.dstream.DStream[(String, Int)] = org.apache.spark.streaming.dstream.MappedDStream@2f1
52472

scala>

scala> val initialRDD = ssc.sparkContext.parallelize(List[(String,Int)]())
initialRDD: org.apache.spark.rdd.RDD[(String, Int)] = ParallelCollectionRDD[0] at parallelize at <console>:38

scala>

scala> val mappingFunc = (word : String , one : Option[Int] , state : State[Int]) =>{
     |
     |    val sum = one.getOrElse(0) + state.getOption.getOrElse(0)
     |
     |    val output = (word,sum)
     |
     |    state.update(sum)
     |
     |    output}
mappingFunc: (String, Option[Int], org.apache.spark.streaming.State[Int]) => (String, Int) = <function3>

scala>

scala> val stateDstream = wordDstream.mapWithState(StateSpec.function(mappingFunc).initialState(initialRDD))
stateDstream: org.apache.spark.streaming.dstream.MapWithStateDStream[String,Int,Int,(String, Int)] = org.apache.spark.streami
ng.dstream.MapWithStateDStreamImpl@5552f9e0

scala>

scala> stateDstream.print()

scala>
```

Output:

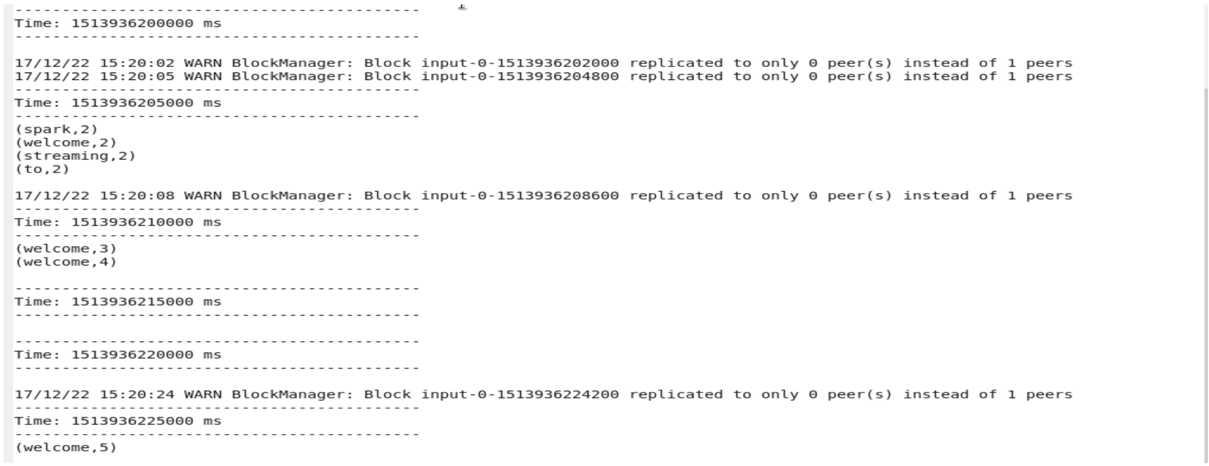
Waiting for input from netcat for stateful streaming , showing checkpoints for welcome, spark, to



```
[acadgild@localhost ~]$ nc -lk 9999
welcome to spark streaming
welcome to spark streaming
welcome
welcome
welcome
```

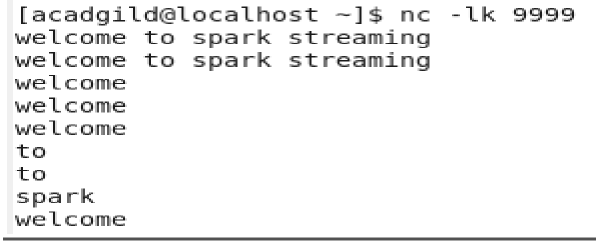


```
scala> ssc.start()

scala>

-------------------------------------------
Time: 1513936170000 ms
-------------------------------------------

-------------------------------------------
Time: 1513936175000 ms
-------------------------------------------

-------------------------------------------
Time: 1513936180000 ms
-------------------------------------------

-------------------------------------------
Time: 1513936185000 ms
-------------------------------------------

-------------------------------------------
Time: 1513936190000 ms
-------------------------------------------

17/12/22 15:19:54 WARN BlockManager: Block input-0-1513936193800 replicated to only 0 peer(s) instead of 1 peers
-------------------------------------------
Time: 1513936195000 ms
-------------------------------------------
(spark,1)
(welcome,1)
(streaming,1)
(to,1)
```

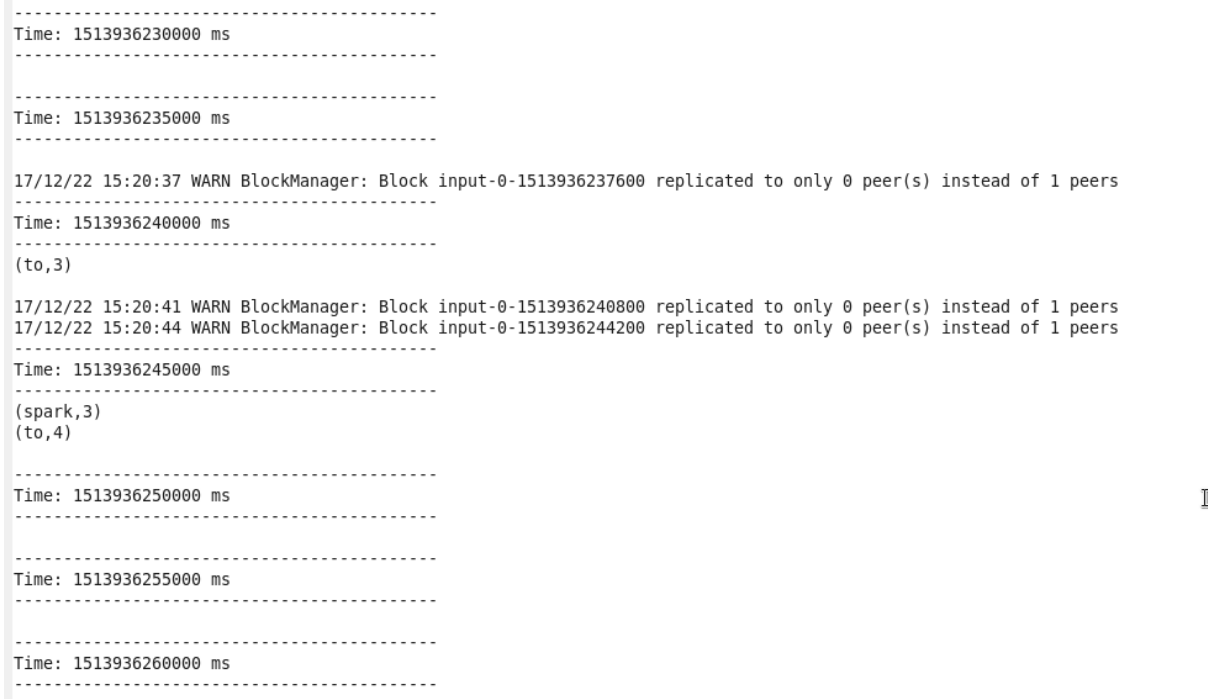


```
-------------------------------------------
Time: 1513936200000 ms
-------------------------------------------

17/12/22 15:20:02 WARN BlockManager: Block input-0-1513936202000 replicated to only 0 peer(s) instead of 1 peers
17/12/22 15:20:05 WARN BlockManager: Block input-0-1513936204800 replicated to only 0 peer(s) instead of 1 peers
-------------------------------------------
Time: 1513936205000 ms
-------------------------------------------
(spark,2)
(welcome,2)
(streaming,2)
(to,2)

17/12/22 15:20:08 WARN BlockManager: Block input-0-1513936208600 replicated to only 0 peer(s) instead of 1 peers
-------------------------------------------
Time: 1513936210000 ms
-------------------------------------------
(welcome,3)
(welcome,4)

-------------------------------------------
Time: 1513936215000 ms
-------------------------------------------

-------------------------------------------
Time: 1513936220000 ms
-------------------------------------------

17/12/22 15:20:24 WARN BlockManager: Block input-0-1513936224200 replicated to only 0 peer(s) instead of 1 peers
-------------------------------------------
Time: 1513936225000 ms
-------------------------------------------
(welcome,5)
```

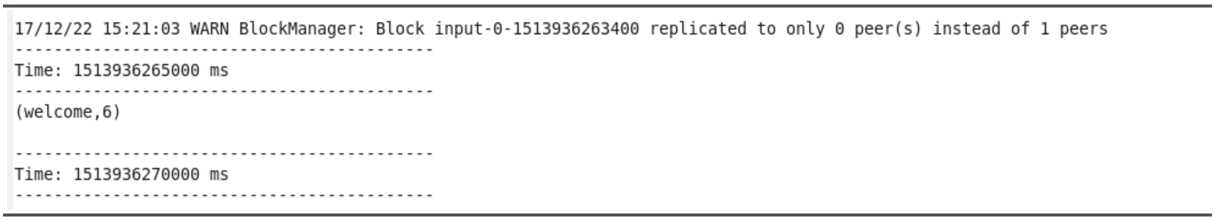
Again giving the input for checkpoints of welcome



```
[acadgild@localhost ~]$ nc -lk 9999
welcome to spark streaming
welcome to spark streaming
welcome
welcome
welcome
to
to
spark
welcome
```



```
-------------------------------------------
Time: 1513936230000 ms
-------------------------------------------

-------------------------------------------
Time: 1513936235000 ms
-------------------------------------------

17/12/22 15:20:37 WARN BlockManager: Block input-0-1513936237600 replicated to only 0 peer(s) instead of 1 peers
-------------------------------------------
Time: 1513936240000 ms
-------------------------------------------
(to,3)

17/12/22 15:20:41 WARN BlockManager: Block input-0-1513936240800 replicated to only 0 peer(s) instead of 1 peers
17/12/22 15:20:44 WARN BlockManager: Block input-0-1513936244200 replicated to only 0 peer(s) instead of 1 peers
-------------------------------------------
Time: 1513936245000 ms
-------------------------------------------
(spark,3)
(to,4)

-------------------------------------------
Time: 1513936250000 ms
-------------------------------------------

-------------------------------------------
Time: 1513936255000 ms
-------------------------------------------

-------------------------------------------
Time: 1513936260000 ms
-------------------------------------------
```



```
17/12/22 15:21:03 WARN BlockManager: Block input-0-1513936263400 replicated to only 0 peer(s) instead of 1 peers
-------------------------------------------
Time: 1513936265000 ms
-------------------------------------------
(welcome,6)

-------------------------------------------
Time: 1513936270000 ms
-------------------------------------------
```