Spark Streaming - TCP

Workflow

1. Start the TPC Server
2. Start Spark Streaming

Find the lab here

Open your note book to 01-Stream-TCP

- Follow along …..
- Create a Spark Context and a Streaming Context

```
from pyspark import SparkContext
from pyspark.streaming import StreamingContext
```

```
sc = SparkContext(master="local[2]", appName="Test Spark Streaming")
ssc = StreamingContext(sc, 1)
ssc.checkpoint("checkpoint")
```

```
ssc = new StreamingContext(sc, 1)
```

The second parameter of (sc, 1), represents the time interval at which streaming data will be divided into batches.

After a StreamingContext is defined you do the followin:

- Define the input sources
- Applying transformations on the Dstreams
- Define output operations on DStreams.
- Start receiving data – streamingContext.start()
- Stop processing streamingContext.stop()

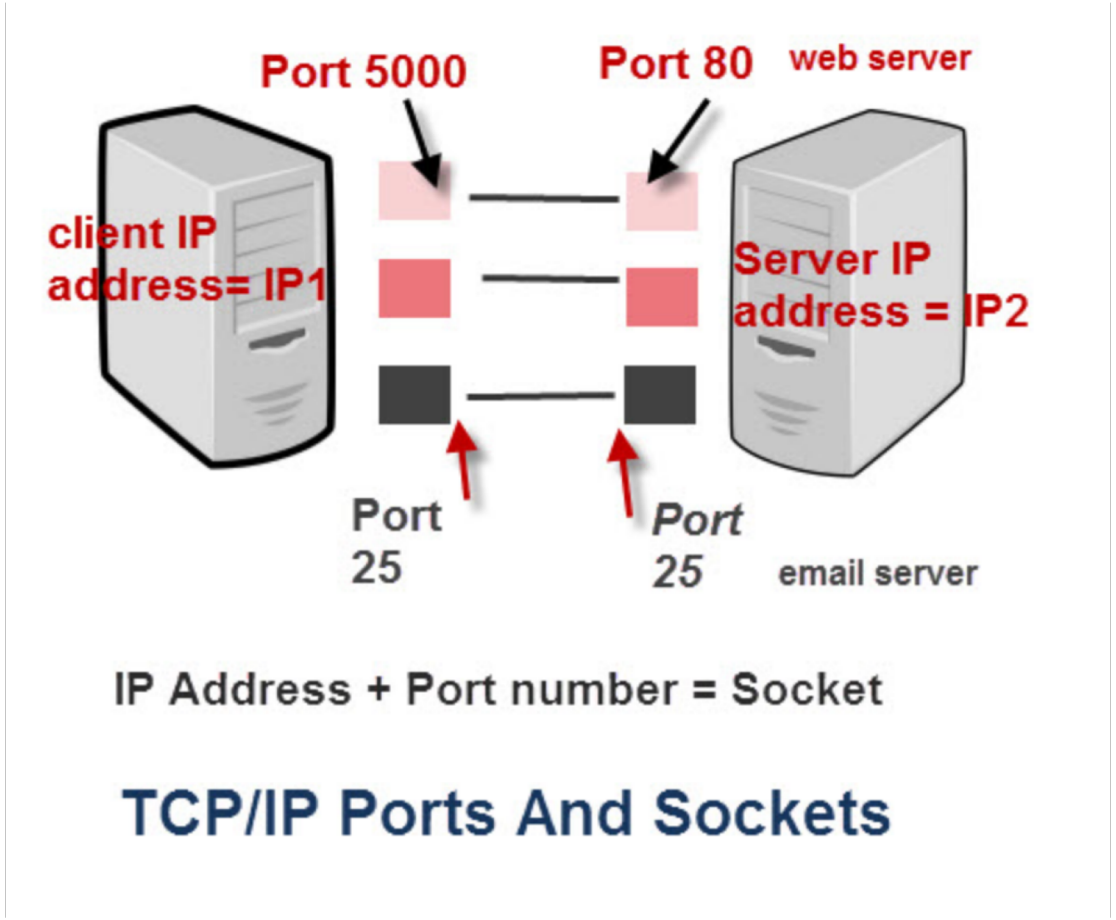
Define Transformations

```
lines = ssc.socketTextStream("nc", 5555)
```

```
words = lines.flatMap(lambda line: line.split(" "))

# Count each word in each batch
pairs = words.map(lambda word: (word, 1))
wordCounts = pairs.reduceByKey(lambda x, y: x + y)
# Print the first ten elements of each RDD generated in this DStream to the console
wordCounts.pprint()
```

What is TCP? [source]



Start Streaming

```
ssc.start()
```

```
-------------------------------------------
Time: 2018-12-31 00:10:10
-------------------------------------------

-------------------------------------------
Time: 2018-12-31 00:10:11
-------------------------------------------

-------------------------------------------
Time: 2018-12-31 00:10:12
-------------------------------------------

-------------------------------------------
Time: 2018-12-31 00:10:13
-------------------------------------------

-------------------------------------------
Time: 2018-12-31 00:10:14
-------------------------------------------
```

```
ssc.stop(stopSparkContext=True, stopGraceFully=True)
```

Tear down the lab

- Follow instructions to terminate the NC server and Streaming container