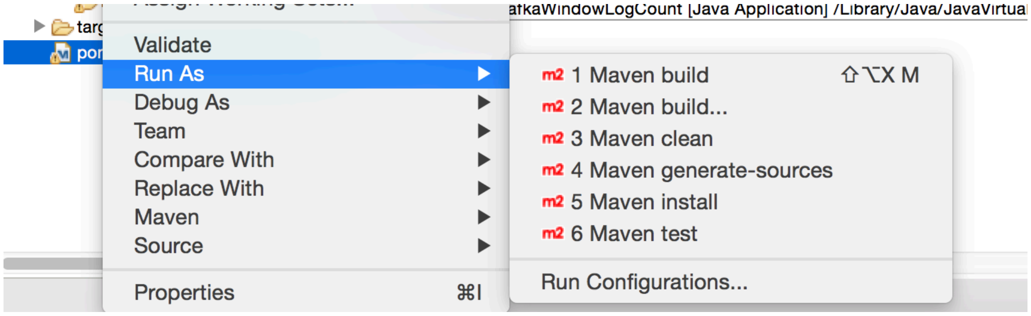
# P5 Spark Kafka Log Aggregation on VM

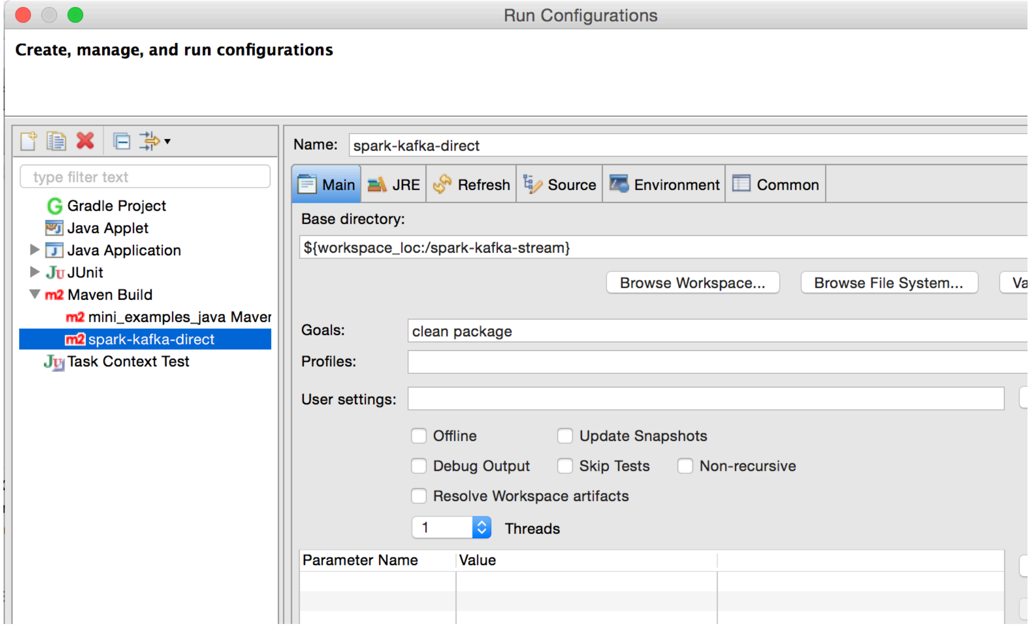
VM: CentOS_6.7_CDH5.5_Spark - CentOS6.7 + Spark1.62 + JDK1.8

**Running out KafkaWindowLogCount app on VM**

Create a JAR via Maven 'package':
Right-click on the pom.xml and select 'Run As' -> 'Run Configurations ...'



Double-click on the 'Maven Build' to create a new configuration - 'spark-kafka-direct' below:



Name: spark-kafka-direct

Base directory:
${workspace_loc:/spark-kafka-stream}

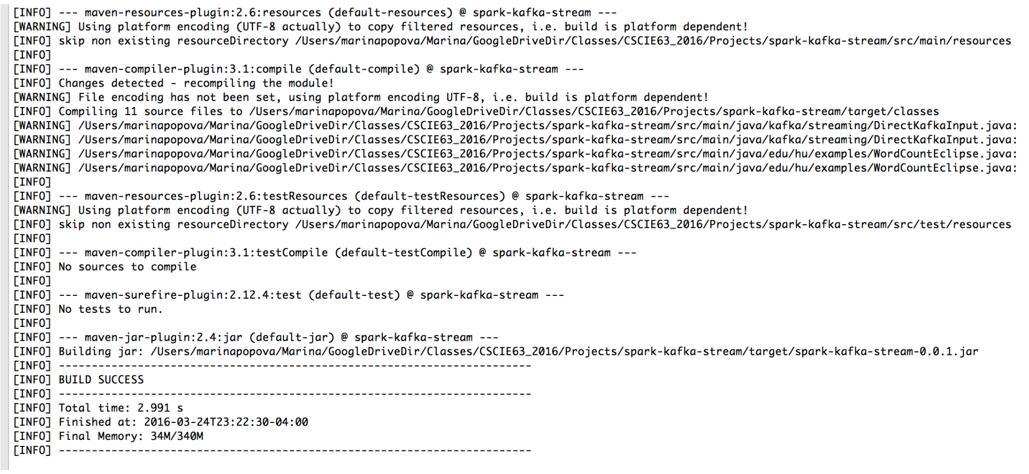
Goals: clean package

Profiles:

User settings:

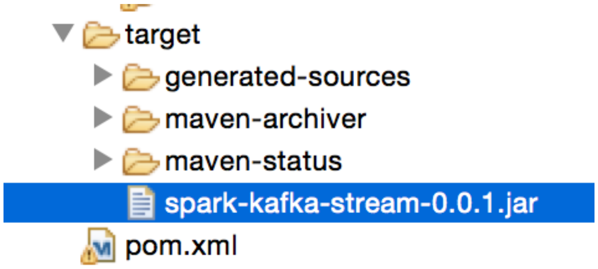
□ Offline □ Update Snapshots
□ Debug Output □ Skip Tests □ Non-recursive
□ Resolve Workspace artifacts
1 Threads

Specify Goals: "clean package"
Run



```
[INFO] --- maven-resources-plugin:2.6:resources (default-resources) @ spark-kafka-stream ---
[WARNING] Using platform encoding (UTF-8 actually) to copy filtered resources, i.e. build is platform dependent!
[INFO] skip non existing resourceDirectory /Users/marinapopova/Marina/GoogleDriveDir/Classes/CSCIE63_2016/Projects/spark-kafka-stream/src/main/resources
[INFO]
[INFO] --- maven-compiler-plugin:3.1:compile (default-compile) @ spark-kafka-stream ---
[INFO] Changes detected - recompiling the module!
[WARNING] File encoding has not been set, using platform encoding UTF-8, i.e. build is platform dependent!
[INFO] Compiling 11 source files to /Users/marinapopova/Marina/GoogleDriveDir/Classes/CSCIE63_2016/Projects/spark-kafka-stream/target/classes
[WARNING] /Users/marinapopova/Marina/GoogleDriveDir/Classes/CSCIE63_2016/Projects/spark-kafka-stream/src/main/java/kafka/streaming/DirectKafkaInput.java:
[WARNING] /Users/marinapopova/Marina/GoogleDriveDir/Classes/CSCIE63_2016/Projects/spark-kafka-stream/src/main/java/kafka/streaming/DirectKafkaInput.java:
[WARNING] /Users/marinapopova/Marina/GoogleDriveDir/Classes/CSCIE63_2016/Projects/spark-kafka-stream/src/main/java/edu/hu/examples/WordCountEclipse.java:
[WARNING] /Users/marinapopova/Marina/GoogleDriveDir/Classes/CSCIE63_2016/Projects/spark-kafka-stream/src/main/java/edu/hu/examples/WordCountEclipse.java:
[INFO]
[INFO] --- maven-resources-plugin:2.6:testResources (default-testResources) @ spark-kafka-stream ---
[WARNING] Using platform encoding (UTF-8 actually) to copy filtered resources, i.e. build is platform dependent!
[INFO] skip non existing resourceDirectory /Users/marinapopova/Marina/GoogleDriveDir/Classes/CSCIE63_2016/Projects/spark-kafka-stream/src/test/resources
[INFO]
[INFO] --- maven-compiler-plugin:3.1:testCompile (default-testCompile) @ spark-kafka-stream ---
[INFO] No sources to compile
[INFO]
[INFO] --- maven-surefire-plugin:2.12.4:test (default-test) @ spark-kafka-stream ---
[INFO] No tests to run.
[INFO]
[INFO] --- maven-jar-plugin:2.4:jar (default-jar) @ spark-kafka-stream ---
[INFO] Building jar: /Users/marinapopova/Marina/GoogleDriveDir/Classes/CSCIE63_2016/Projects/spark-kafka-stream/target/spark-kafka-stream-0.0.1.jar
[INFO] ------------------------------------------------------------------------
[INFO] BUILD SUCCESS
[INFO] ------------------------------------------------------------------------
[INFO] Total time: 2.991 s
[INFO] Finished at: 2016-03-24T23:22:30-04:00
[INFO] Final Memory: 34M/340M
[INFO] ------------------------------------------------------------------------
```

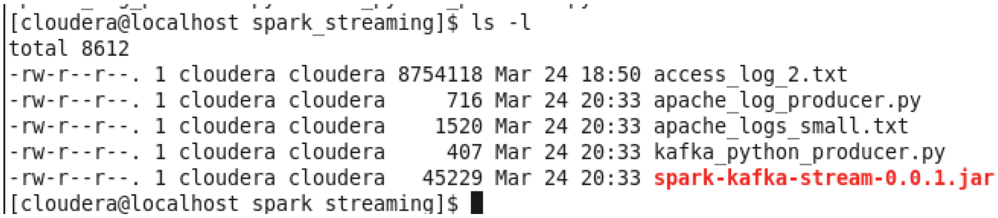
Created artifact:



```
▼ target
    ▶ generated-sources
    ▶ maven-archiver
    ▶ maven-status
      spark-kafka-stream-0.0.1.jar
  pom.xml
```

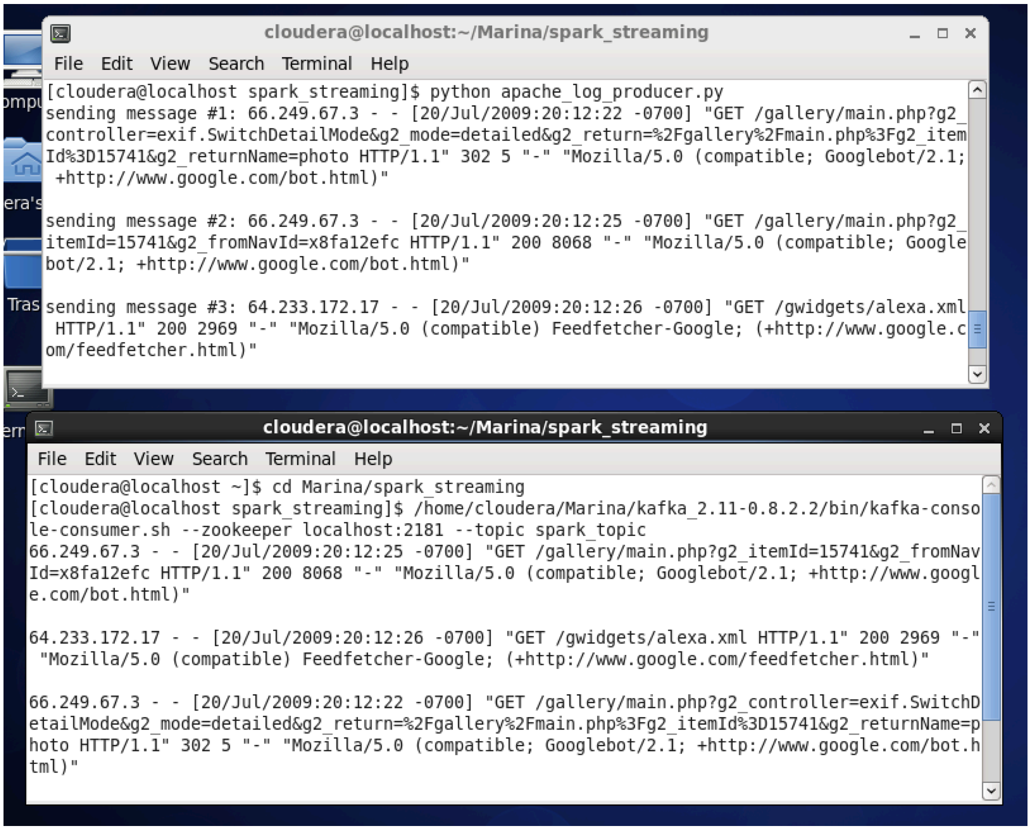
**NOTE: You might as well just run 'mvn clean package' from a command line on either your local laptop or your VM.**

Copy Spark job jar, Python scripts and log files to VM:



```
[cloudera@localhost spark_streaming]$ ls -l
total 8612
-rw-r--r--. 1 cloudera cloudera 8754118 Mar 24 18:50 access_log_2.txt
-rw-r--r--. 1 cloudera cloudera     716 Mar 24 20:33 apache_log_producer.py
-rw-r--r--. 1 cloudera cloudera    1520 Mar 24 20:33 apache_logs_small.txt
-rw-r--r--. 1 cloudera cloudera     407 Mar 24 20:33 kafka_python_producer.py
-rw-r--r--. 1 cloudera cloudera   45229 Mar 24 20:33 spark-kafka-stream-0.0.1.jar
[cloudera@localhost spark_streaming]$ █
```

Start kafka console consumer to test the python log producer:
/home/cloudera/Marina/kafka_2.11-0.8.2.2/**bin/kafka-console-consumer.sh --zookeeper localhost:2181 --topic** spark_topic


cloudera@localhost:~/Marina/spark_streaming
File Edit View Search Terminal Help

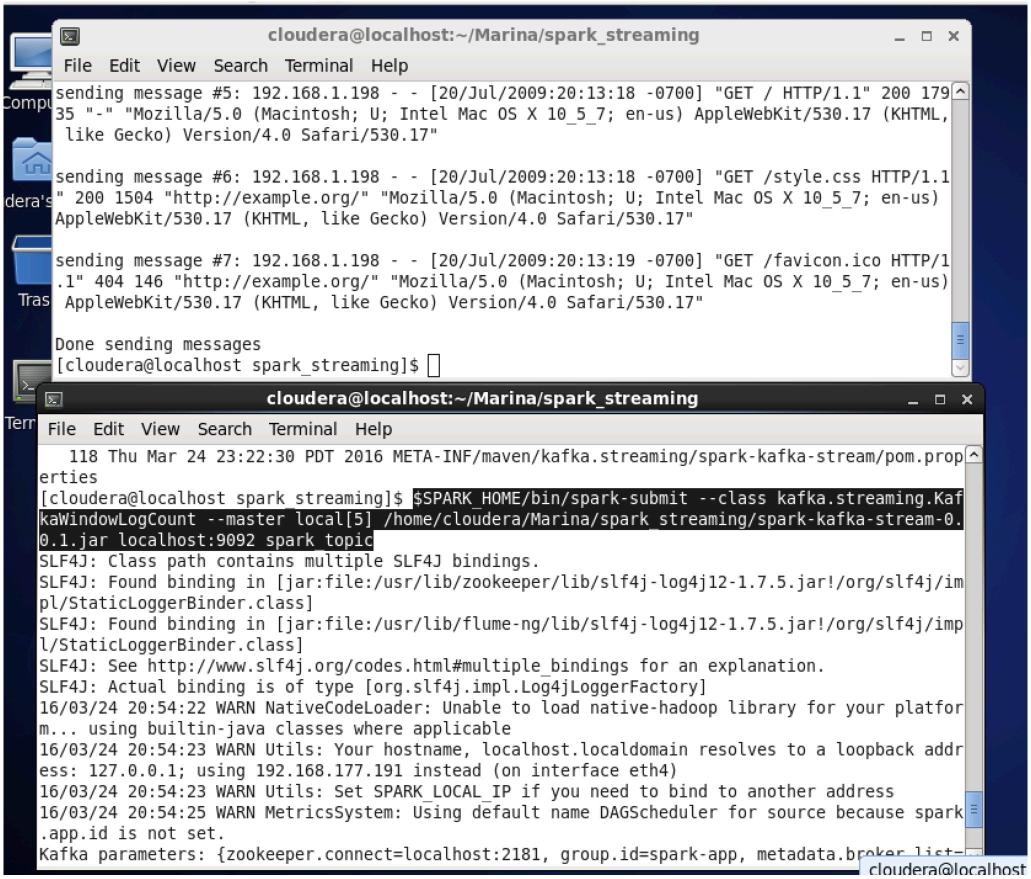
```
[cloudera@localhost spark_streaming]$ python apache_log_producer.py
sending message #1: 66.249.67.3 - - [20/Jul/2009:20:12:22 -0700] "GET /gallery/main.php?g2_
controller=exif.SwitchDetailMode&g2_mode=detailed&g2_return=%2Fgallery%2Fmain.php%3Fg2_item
Id%3D15741&g2_returnName=photo HTTP/1.1" 302 5 "-" "Mozilla/5.0 (compatible; Googlebot/2.1;
 +http://www.google.com/bot.html)"

sending message #2: 66.249.67.3 - - [20/Jul/2009:20:12:25 -0700] "GET /gallery/main.php?g2_
itemId=15741&g2_fromNavId=x8fa12efc HTTP/1.1" 200 8068 "-" "Mozilla/5.0 (compatible; Google
bot/2.1; +http://www.google.com/bot.html)"

sending message #3: 64.233.172.17 - - [20/Jul/2009:20:12:26 -0700] "GET /gwidgets/alexa.xml
 HTTP/1.1" 200 2969 "-" "Mozilla/5.0 (compatible) Feedfetcher-Google; (+http://www.google.c
om/feedfetcher.html)"
```

cloudera@localhost:~/Marina/spark_streaming
File Edit View Search Terminal Help

```
[cloudera@localhost ~]$ cd Marina/spark_streaming
[cloudera@localhost spark_streaming]$ /home/cloudera/Marina/kafka_2.11-0.8.2.2/bin/kafka-conso
le-consumer.sh --zookeeper localhost:2181 --topic spark_topic
66.249.67.3 - - [20/Jul/2009:20:12:25 -0700] "GET /gallery/main.php?g2_itemId=15741&g2_fromNav
Id=x8fa12efc HTTP/1.1" 200 8068 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.googl
e.com/bot.html)"

64.233.172.17 - - [20/Jul/2009:20:12:26 -0700] "GET /gwidgets/alexa.xml HTTP/1.1" 200 2969 "-"
 "Mozilla/5.0 (compatible) Feedfetcher-Google; (+http://www.google.com/feedfetcher.html)"

66.249.67.3 - - [20/Jul/2009:20:12:22 -0700] "GET /gallery/main.php?g2_controller=exif.SwitchD
etailMode&g2_mode=detailed&g2_return=%2Fgallery%2Fmain.php%3Fg2_itemId%3D15741&g2_returnName=p
hoto HTTP/1.1" 302 5 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.h
tml)"
```

Now run the KafkaWindowLogCount job:

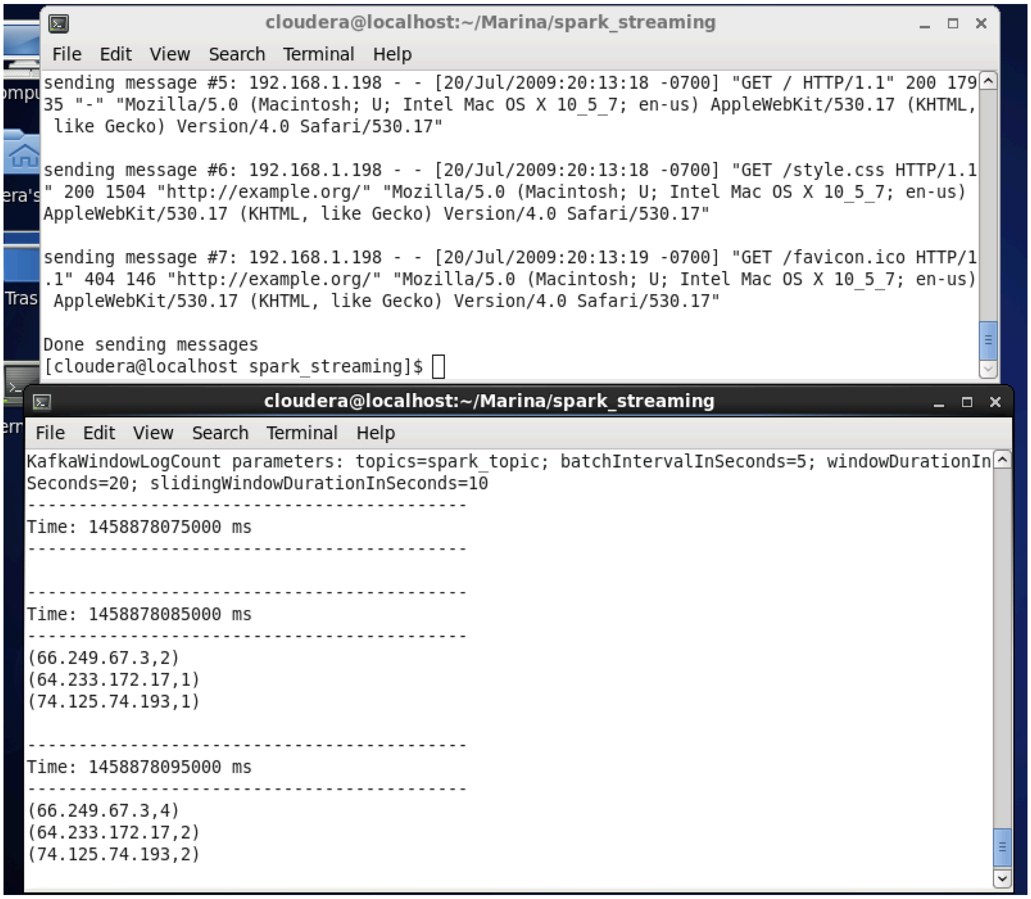
**$SPARK_HOME/bin/spark-submit --class kafka.streaming.KafkaWindowLogCount --master local[5] /home/cloudera/Marina/spark_streaming/spark-kafka-stream-0.0.1.jar localhost:9092 spark_topic 5 20 10**


cloudera@localhost:~/Marina/spark_streaming
File Edit View Search Terminal Help
Compu
dera's
Tras
Terr
cloudera@localhost:~/Marina/spark_streaming
File Edit View Search Terminal Help
cloudera@localhost

```
sending message #5: 192.168.1.198 - - [20/Jul/2009:20:13:18 -0700] "GET / HTTP/1.1" 200 179
35 "-" "Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_5_7; en-us) AppleWebKit/530.17 (KHTML,
 like Gecko) Version/4.0 Safari/530.17"

sending message #6: 192.168.1.198 - - [20/Jul/2009:20:13:18 -0700] "GET /style.css HTTP/1.1
" 200 1504 "http://example.org/" "Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_5_7; en-us)
AppleWebKit/530.17 (KHTML, like Gecko) Version/4.0 Safari/530.17"

sending message #7: 192.168.1.198 - - [20/Jul/2009:20:13:19 -0700] "GET /favicon.ico HTTP/1
.1" 404 146 "http://example.org/" "Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_5_7; en-us)
 AppleWebKit/530.17 (KHTML, like Gecko) Version/4.0 Safari/530.17"

Done sending messages
[cloudera@localhost spark_streaming]$ 
```

```
  118 Thu Mar 24 23:22:30 PDT 2016 META-INF/maven/kafka.streaming/spark-kafka-stream/pom.prop
erties
[cloudera@localhost spark_streaming]$ $SPARK_HOME/bin/spark-submit --class kafka.streaming.Kaf
kaWindowLogCount --master local[5] /home/cloudera/Marina/spark_streaming/spark-kafka-stream-0.
0.1.jar localhost:9092 spark_topic
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/zookeeper/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/im
pl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/flume-ng/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/imp
l/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
16/03/24 20:54:22 WARN NativeCodeLoader: Unable to load native-hadoop library for your platfor
m... using builtin-java classes where applicable
16/03/24 20:54:23 WARN Utils: Your hostname, localhost.localdomain resolves to a loopback addr
ess: 127.0.0.1; using 192.168.177.191 instead (on interface eth4)
16/03/24 20:54:23 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
16/03/24 20:54:25 WARN MetricsSystem: Using default name DAGScheduler for source because spark
.app.id is not set.
Kafka parameters: {zookeeper.connect=localhost:2181, group.id=spark-app, metadata.broker.list=
```

Output from the spark app:

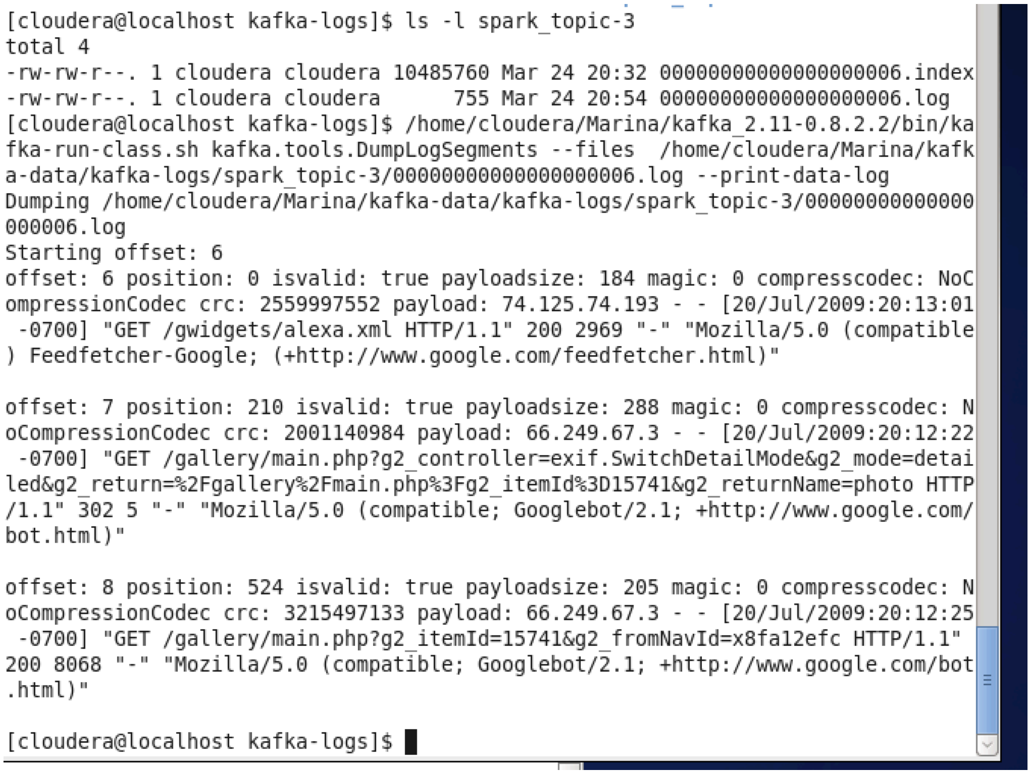

cloudera@localhost:~/Marina/spark_streaming
File Edit View Search Terminal Help

```
sending message #5: 192.168.1.198 - - [20/Jul/2009:20:13:18 -0700] "GET / HTTP/1.1" 200 179
35 "-" "Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_5_7; en-us) AppleWebKit/530.17 (KHTML,
 like Gecko) Version/4.0 Safari/530.17"

sending message #6: 192.168.1.198 - - [20/Jul/2009:20:13:18 -0700] "GET /style.css HTTP/1.1
" 200 1504 "http://example.org/" "Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_5_7; en-us)
AppleWebKit/530.17 (KHTML, like Gecko) Version/4.0 Safari/530.17"

sending message #7: 192.168.1.198 - - [20/Jul/2009:20:13:19 -0700] "GET /favicon.ico HTTP/1
.1" 404 146 "http://example.org/" "Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_5_7; en-us)
 AppleWebKit/530.17 (KHTML, like Gecko) Version/4.0 Safari/530.17"

Done sending messages
[cloudera@localhost spark_streaming]$ 
```

cloudera@localhost:~/Marina/spark_streaming
File Edit View Search Terminal Help

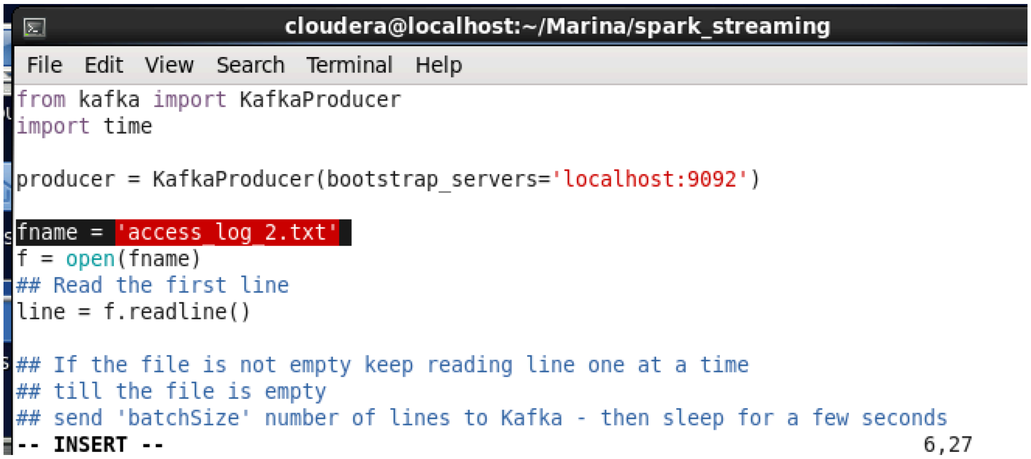
```
KafkaWindowLogCount parameters: topics=spark_topic; batchIntervalInSeconds=5; windowDurationIn
Seconds=20; slidingWindowDurationInSeconds=10
-------------------------------------------
Time: 1458878075000 ms
-------------------------------------------

-------------------------------------------
Time: 1458878085000 ms
-------------------------------------------
(66.249.67.3,2)
(64.233.172.17,1)
(74.125.74.193,1)

-------------------------------------------
Time: 1458878095000 ms
-------------------------------------------
(66.249.67.3,4)
(64.233.172.17,2)
(74.125.74.193,2)
```

Examine Kafka logs:



```
[cloudera@localhost kafka-logs]$ ls -l spark_topic-3
total 4
-rw-rw-r--. 1 cloudera cloudera 10485760 Mar 24 20:32 00000000000000000006.index
-rw-rw-r--. 1 cloudera cloudera      755 Mar 24 20:54 00000000000000000006.log
[cloudera@localhost kafka-logs]$ /home/cloudera/Marina/kafka_2.11-0.8.2.2/bin/ka
fka-run-class.sh kafka.tools.DumpLogSegments --files  /home/cloudera/Marina/kafk
a-data/kafka-logs/spark_topic-3/00000000000000000006.log --print-data-log
Dumping /home/cloudera/Marina/kafka-data/kafka-logs/spark_topic-3/00000000000000
000006.log
Starting offset: 6
offset: 6 position: 0 isvalid: true payloadsize: 184 magic: 0 compresscodec: NoC
ompressionCodec crc: 2559997552 payload: 74.125.74.193 - - [20/Jul/2009:20:13:01
 -0700] "GET /gwidgets/alexa.xml HTTP/1.1" 200 2969 "-" "Mozilla/5.0 (compatible
) Feedfetcher-Google; (+http://www.google.com/feedfetcher.html)"

offset: 7 position: 210 isvalid: true payloadsize: 288 magic: 0 compresscodec: N
oCompressionCodec crc: 2001140984 payload: 66.249.67.3 - - [20/Jul/2009:20:12:22
 -0700] "GET /gallery/main.php?g2_controller=exif.SwitchDetailMode&g2_mode=detai
led&g2_return=%2Fgallery%2Fmain.php%3Fg2_itemId%3D15741&g2_returnName=photo HTTP
/1.1" 302 5 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/
bot.html)"

offset: 8 position: 524 isvalid: true payloadsize: 205 magic: 0 compresscodec: N
oCompressionCodec crc: 3215497133 payload: 66.249.67.3 - - [20/Jul/2009:20:12:25
 -0700] "GET /gallery/main.php?g2_itemId=15741&g2_fromNavId=x8fa12efc HTTP/1.1"
200 8068 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot
.html)"

[cloudera@localhost kafka-logs]$ 
```

Processing the large Apache log file:



cloudera@localhost:~/Marina/spark_streaming
File Edit View Search Terminal Help

```
from kafka import KafkaProducer
import time

producer = KafkaProducer(bootstrap_servers='localhost:9092')

fname = 'access_log_2.txt'
f = open(fname)
## Read the first line
line = f.readline()

## If the file is not empty keep reading line one at a time
## till the file is empty
## send 'batchSize' number of lines to Kafka - then sleep for a few seconds
-- INSERT --                                                        6,27
```

Output from the Spark app:

*[cloudera@localhost spark_streaming]$* ***$SPARK_HOME/bin/spark-submit --class kafka.streaming.KafkaWindowLogCount --master local[5] /home/cloudera/Marina/spark_streaming/spark-kafka-stream-0.0.1.jar localhost:9092 spark_topic 10 30 20***

...

```
16/03/24 21:06:11 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
16/03/24 21:06:13 WARN MetricsSystem: Using default name DAGScheduler for source because
spark.app.id is not set.
Kafka parameters: {zookeeper.connect=localhost:2181, group.id=spark-app,
metadata.broker.list=localhost:9092}
KafkaWindowLogCount parameters: topics=spark_topic; batchIntervalInSeconds=10;
windowDurationInSeconds=30; slidingWindowDurationInSeconds=20
-------------------------------------------
Time: 1458878790000 ms
-------------------------------------------
-------------------------------------------
Time: 1458878810000 ms
-------------------------------------------
-------------------------------------------
Time: 1458878830000 ms
-------------------------------------------
-------------------------------------------
Time: 1458878850000 ms
-------------------------------------------
-------------------------------------------
Time: 1458878870000 ms
-------------------------------------------
(66.249.67.3,8)
(66.249.67.87,6)
(72.30.142.87,1)
(65.55.106.132,1)
(65.55.106.160,2)
(220.181.7.30,1)
(64.233.173.2,2)
(64.233.172.17,1)
(192.168.1.198,7)
(74.125.74.193,4)
...
-------------------------------------------
Time: 1458878890000 ms
-------------------------------------------
(66.249.67.3,14)
(66.249.67.87,2)
(72.30.142.87,2)
(74.125.16.65,1)
(67.183.157.181,1)
(72.14.194.1,3)
(220.181.7.13,1)
(64.233.173.2,3)
(74.125.75.17,1)
```

```
(74.15.53.228,1)
...
  ^C[cloudera@localhost spark_streaming]$
```