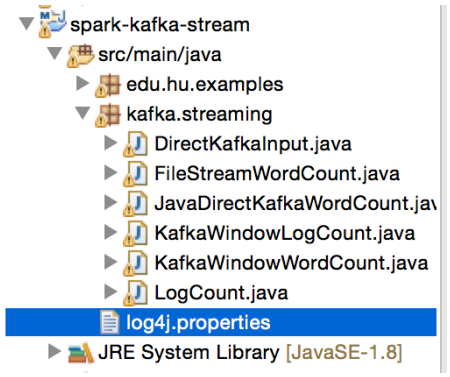
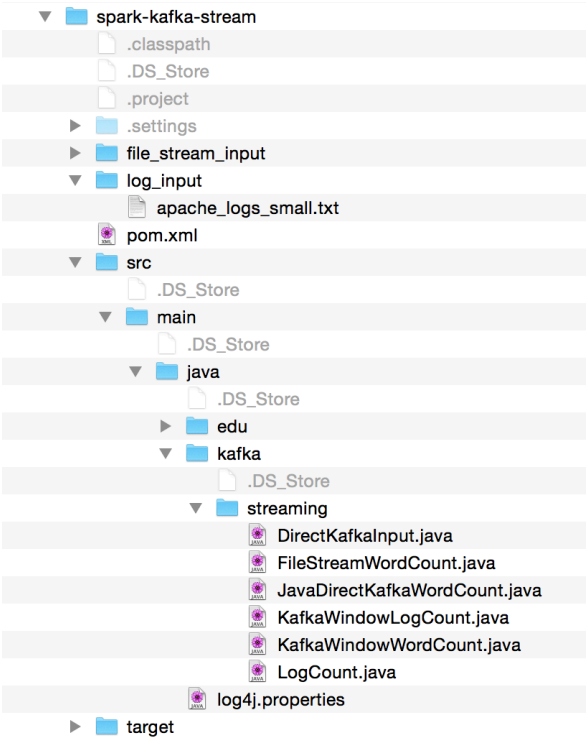
P4 Spark Kafka local (Eclipse) setup

Create a regular Maven project folder structure , with pom.xml
Add log4j.properties to your src/main/java dir (if you want to control logging levels) :



An example of my project "spark-kafka-stream" structure:

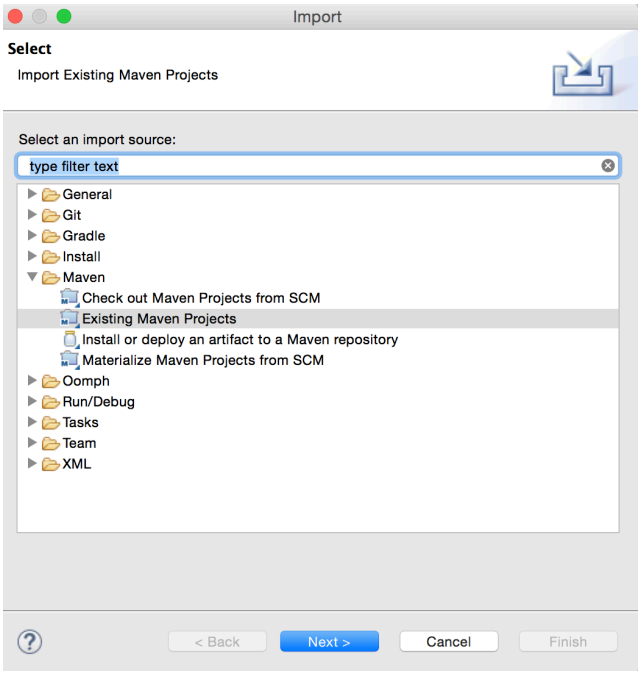


Most important pom.xml settings:

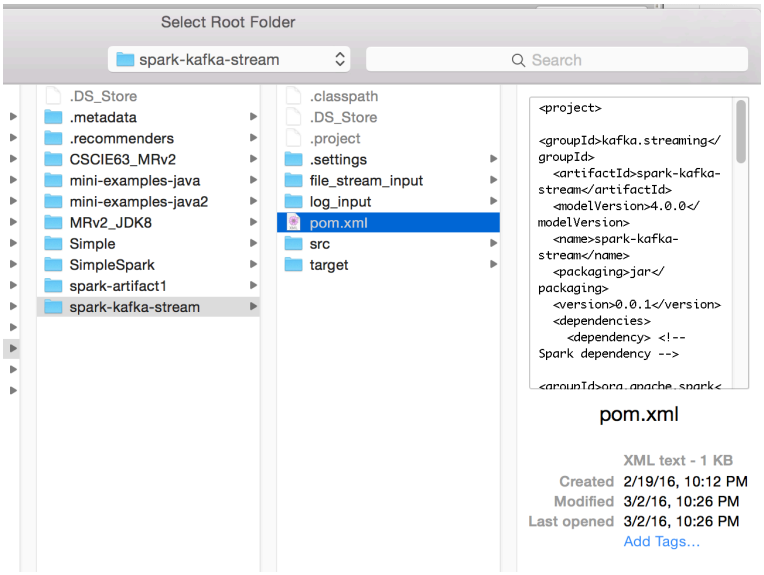
```
<dependencies>
   <dependency> <!-- Spark dependency -->
     <groupId>org.apache.spark</groupId>
```

```
    <artifactId>spark-core_2.10</artifactId>
    <version>1.6.0</version>
    <scope>provided</scope>
  </dependency>
  <dependency>
      <groupId>org.apache.spark</groupId>
      <artifactId>spark-streaming_2.10</artifactId>
      <version>1.6.0</version>
      <scope>provided</scope>
  </dependency>
  <dependency>
      <groupId>org.apache.spark</groupId>
      <artifactId>spark-streaming-kafka_2.10</artifactId>
      <version>1.6.0</version>
  </dependency>
</dependencies>
<properties>
  <java.version>1.8</java.version>
</properties>
```

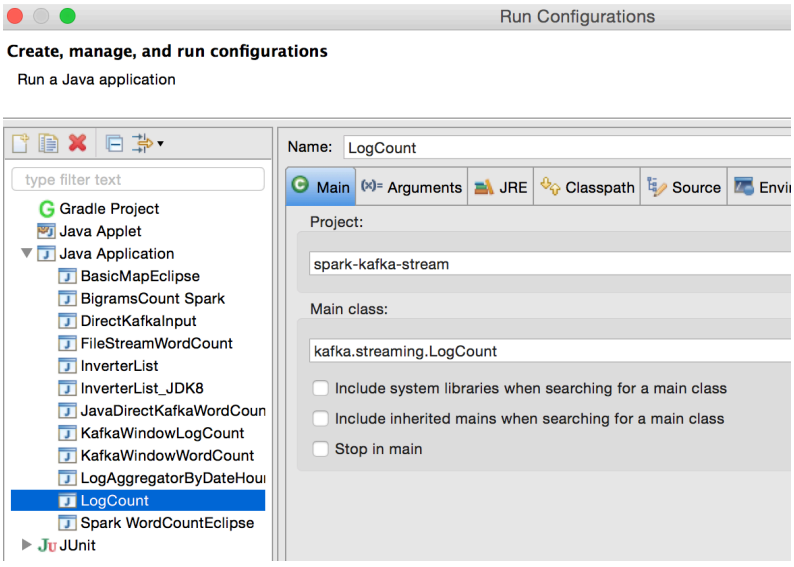
Create a new Java Maven project in Eclipse by doing "File --> Import... --> Maven -> Existing Maven project ..."



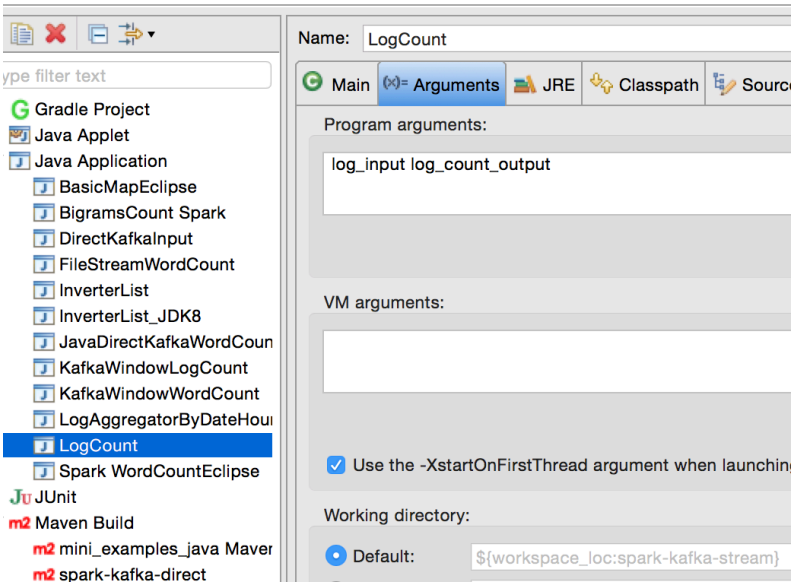
Navigate to your project's directory:



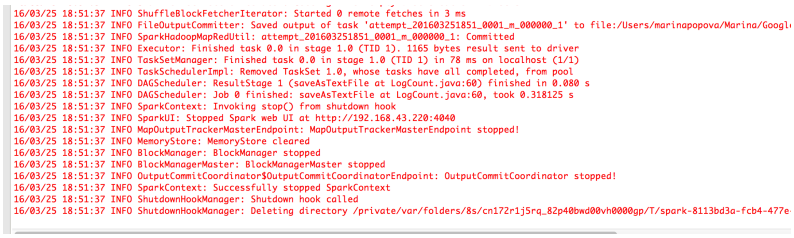
First, lets try to run a simple Spark streaming class that reads Apache logs from a file, and aggregates logs by their client Ips.
We will use the same Regex pattern that you used in the Hive data import - to parse the standard Apache log format.
Class: LogCount.java
To run - create a new Run Configuration:



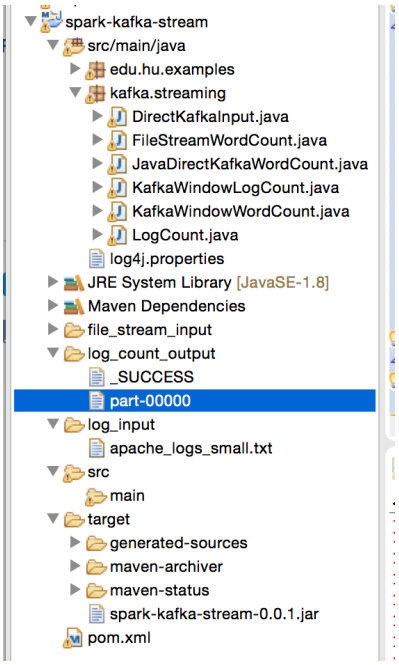
Set input parameters - input apache logs dir and output dir for the results (make sure it does NOT exist):



Run:

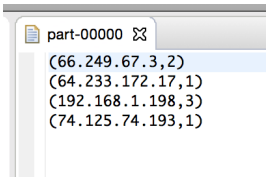


Results are in the specified dir:



We used apache_logs_small.txt as input:

Results of the LogCount:



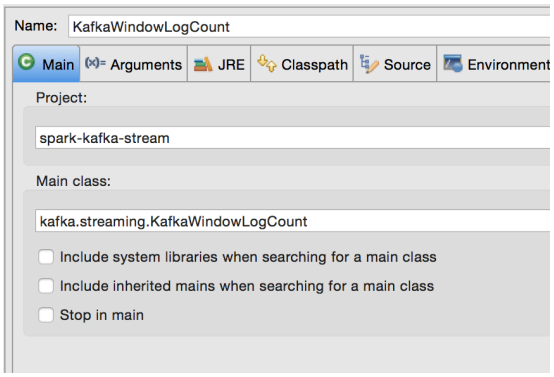
As you can see - the counts are correct!

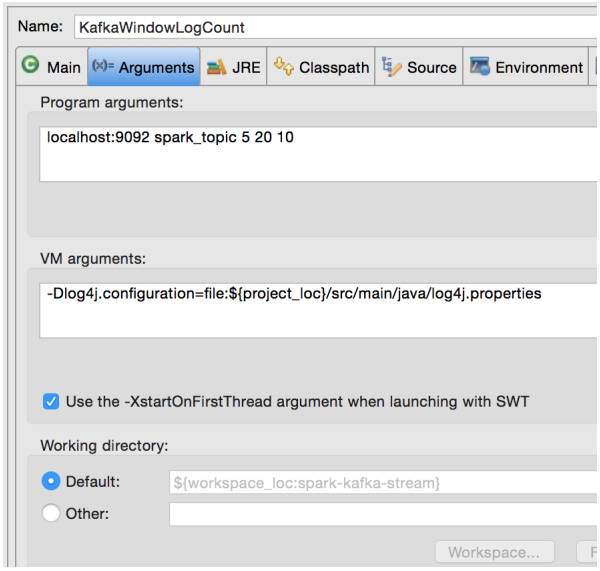
Now we can finally run the Spark streaming app that reads logs from Kafka and aggregates them over specified windows by clientIPs:

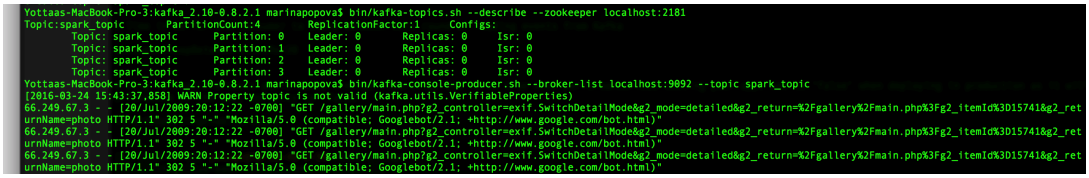
Java class: KafkaWindowLogCount.java

First, make sure your Kafka (and Zookeeper) are running.

Create a Run Configuration for the KafkaWindowLogCount:







Name: KafkaWindowLogCount

Project:

spark-kafka-stream

Main class:

kafka.streaming.KafkaWindowLogCount

☐ Include system libraries when searching for a main class

☐ Include inherited mains when searching for a main class

☐ Stop in main

Name: KafkaWindowLogCount

Program arguments:

localhost:9092 spark_topic 5 20 10

VM arguments:

-Dlog4j.configuration=file:${project_loc}/src/main/java/log4j.properties

☑ Use the -XstartOnFirstThread argument when launching with SWT

Working directory:

● Default: ${workspace_loc:spark-kafka-stream}

○ Other:

Run the application

We will test with two methods of sending logs to Kafka:

1. Using kafka-console-producer - send a few logs manually (same IP):

```
Yottaas-MacBook-Pro-3:kafka_2.10-0.8.2.1 marinapopova$ bin/kafka-topics.sh --describe --zookeeper localhost:2181
Topic:spark_topic	PartitionCount:4	ReplicationFactor:1	Configs:
	Topic: spark_topic	Partition: 0	Leader: 0	Replicas: 0	Isr: 0
	Topic: spark_topic	Partition: 1	Leader: 0	Replicas: 0	Isr: 0
	Topic: spark_topic	Partition: 2	Leader: 0	Replicas: 0	Isr: 0
	Topic: spark_topic	Partition: 3	Leader: 0	Replicas: 0	Isr: 0
Yottaas-MacBook-Pro-3:kafka_2.10-0.8.2.1 marinapopova$ bin/kafka-console-producer.sh --broker-list localhost:9092 --topic spark_topic
[2016-03-24 15:43:37,858] WARN Property topic is not valid (kafka.utils.VerifiableProperties)
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_itemId%3D15741&g2_ret
urnName=photo HTTP/1.1" 302 5 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"
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_itemId%3D15741&g2_ret
urnName=photo HTTP/1.1" 302 5 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"
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_itemId%3D15741&g2_ret
urnName=photo HTTP/1.1" 302 5 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"
```

Notice correct aggregated count by IP in Spark console output: (in Eclipse):

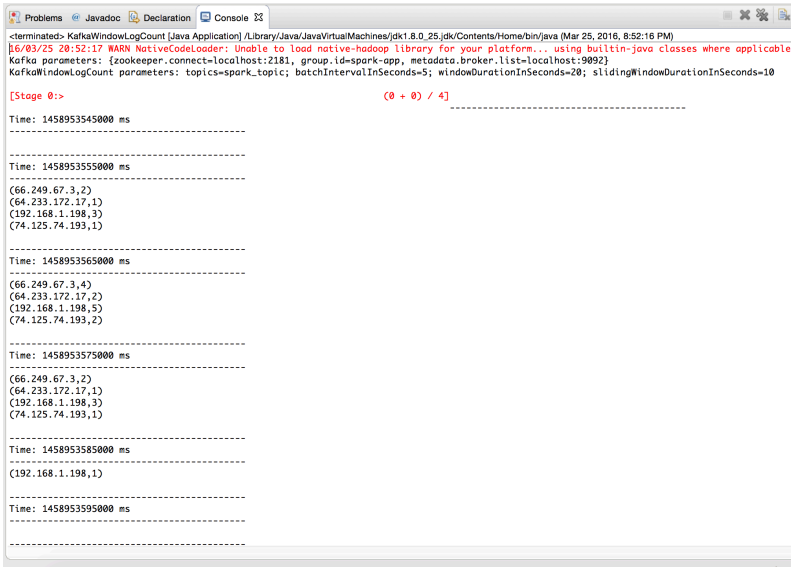
```
16/03/24 15:45:50 INFO TaskSetManager: Finished task 2.0 in stage 43.0 (TID 63) in 4 ms on localhost (4/4)
16/03/24 15:45:50 INFO TaskSchedulerImpl: Removed TaskSet 43.0, whose tasks have all completed, from pool 
16/03/24 15:45:50 INFO DAGScheduler: ResultStage 43 (print at KafkaWindowLogCount.java:145) finished in 0.005 s
16/03/24 15:45:50 INFO DAGScheduler: Job 19 finished: print at KafkaWindowLogCount.java:145, took 0.011499 s
-------------------------------------------
Time: 1458848750000 ms
-------------------------------------------
(66.249.67.3,3)

16/03/24 15:45:50 INFO JobScheduler: Finished job streaming job 1458848750000 ms.1 from job set of time 1458848750000 ms
16/03/24 15:45:50 INFO JobScheduler: Total delay: 0.078 s for time 1458848750000 ms (execution: 0.057 s)
16/03/24 15:45:50 INFO MapPartitionsRDD: Removing RDD 1 from persistence list
```

2. Run the apache log python producer - twice, to get more messages

Yottaas-MBP-3:Section07_KasfkaStreaming marinapopova$ ***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_itemId%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; Googlebot/2.1; +http://www.google.com/bot.html)"
Sleeping for 2 seconds ...
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.com/feedfetcher.html)"
sending message #4: 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)"
Sleeping for 2 seconds ...
sending message #5: 192.168.1.198 - - [20/Jul/2009:20:13:18 -0700] "GET / HTTP/1.1" 200 17935 "-" "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"
Sleeping for 2 seconds ...
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

Yottaas-MBP-3:Section07_KasfkaStreaming marinapopova$ 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_itemId%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; Googlebot/2.1; +http://www.google.com/bot.html)"
Sleeping for 2 seconds ...
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.com/feedfetcher.html)"
sending message #4: 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)"
Sleeping for 2 seconds ...
sending message #5: 192.168.1.198 - - [20/Jul/2009:20:13:18 -0700] "GET / HTTP/1.1" 200 17935 "-" "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"
Sleeping for 2 seconds ...
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
Yottaas-MBP-3:Section07_KasfkaStreaming marinapopova$

See output on the Eclipse console:



16/03/25 20:52:17 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

Kafka parameters: {zookeeper.connect=localhost:2181, group.id=spark-app, metadata.broker.list=localhost:9092}

KafkaWindowLogCount parameters: topics=spark_topic; batchIntervalInSeconds=5; windowDurationInSeconds=20; slidingWindowDurationInSeconds=10

[Stage 0:> (0 + 0) / 4]

Time: 1458953545000 ms <-- start of the sliding window #1

Time: 1458953555000 ms <-- start of the sliding window #2

(66.249.67.3,2)

(64.233.172.17,1)

(192.168.1.198,3)

(74.125.74.193,1)

Time: 1458953565000 ms <-- start of the sliding window #3

(66.249.67.3,4)

(64.233.172.17,2)

(192.168.1.198,5)

(74.125.74.193,2)

Time: 1458953575000 ms

(66.249.67.3,2)

(64.233.172.17,1)

(192.168.1.198,3)

(74.125.74.193,1)

Time: 1458953585000 ms

(192.168.1.198,1)

Time: 1458953595000 ms

Time: 1458953605000 ms
