

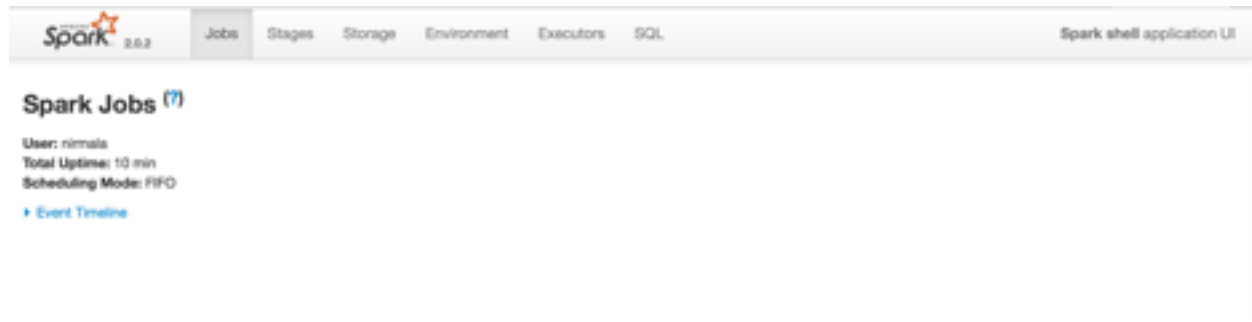
Title: Cloud Mini-HW-4

Spark Streaming

Integrate Spark with Kafka and do some simple processing for streaming data.

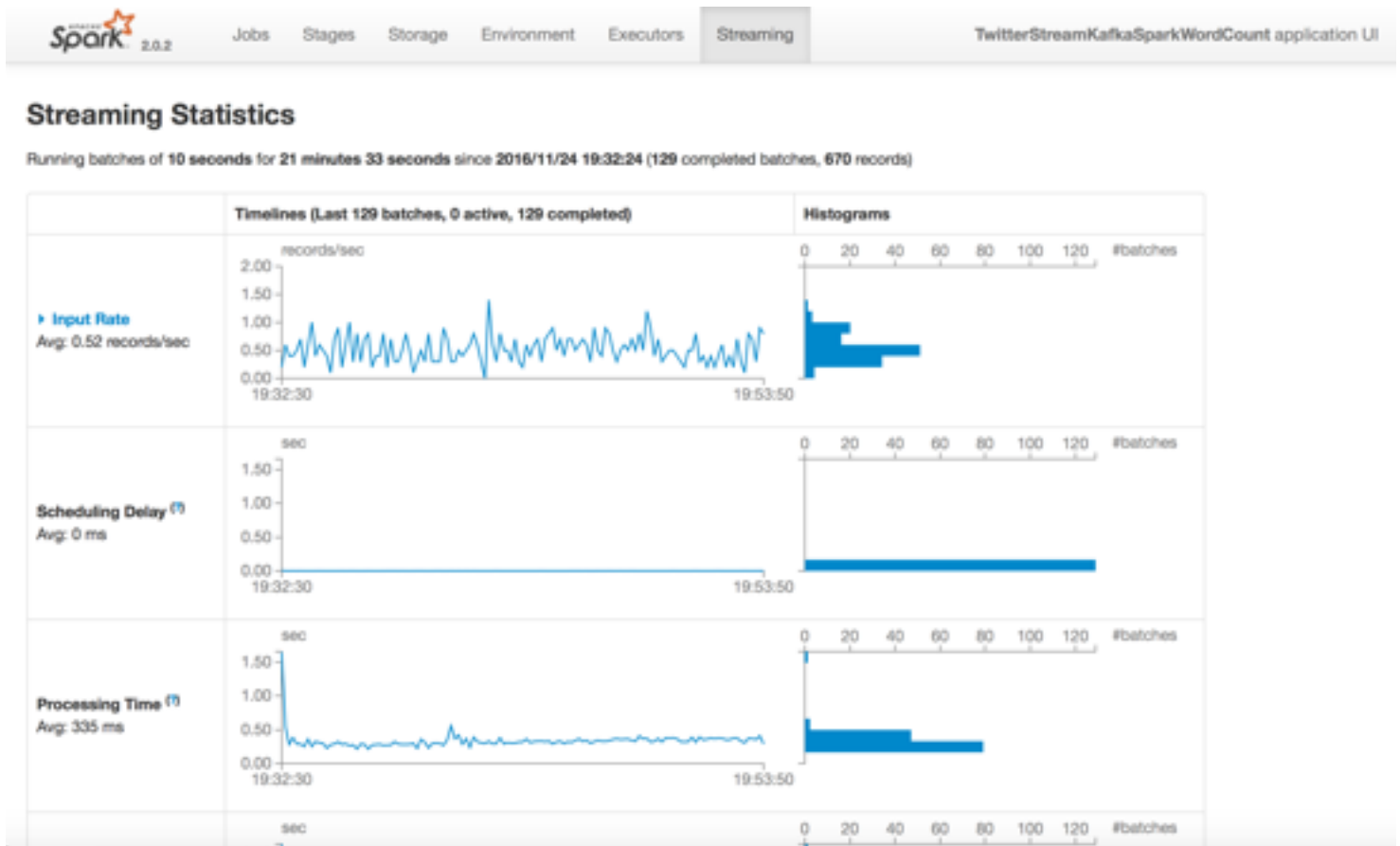
1. The screenshot of Spark UI page when Spark is running(localhost:4040/jobs)

Used pre-built pyspark from Hadoop 2.7 distribution and set up the pyspark environment (local) to run both interactively and with lpython. Below screenshot when pyspark is running and no jobs are submitted.




Thursday, November 24, 2016

Screenshot of Spark UI after submitting WordFrequency job.



Thursday, November 24, 2016

 2.0.2

Jobs

Stages

Storage

Environment

Executors

Streaming

TwitterStreamKafkaSparkWordCount application UI

Spark Jobs ^(?)

User: nirmala
Total Uptime: 20 min
Scheduling Mode: FIFO
Completed Jobs: 352

[Event Timeline](#)

Completed Jobs (352)

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
351	call at /Users/nirmala/spark-2.0.2-bin-hadoop2.7/python/lib/py4j-0.10.3-src.zip/py4j/java_gat...	2016/11/24 19:51:50	30 ms	2/2	<div>9/9</div>
350	runJob at PythonRDD.scala:441	2016/11/24 19:51:50	0.1 s	1/1 (1 skipped)	<div>8/8 (1 skipped)</div>
349	runJob at PythonRDD.scala:441	2016/11/24 19:51:50	40 ms	2/2	<div>2/2</div>
348	call at /Users/nirmala/spark-2.0.2-bin-hadoop2.7/python/lib/py4j-0.10.3-src.zip/py4j/java_gat...	2016/11/24 19:51:40	34 ms	2/2	<div>9/9</div>
347	runJob at PythonRDD.scala:441	2016/11/24 19:51:40	0.1 s	1/1 (1 skipped)	<div>8/8 (1 skipped)</div>
346	runJob at PythonRDD.scala:441	2016/11/24 19:51:40	80 ms	2/2	<div>2/2</div>
345	call at /Users/nirmala/spark-2.0.2-bin-hadoop2.7/python/lib/py4j-0.10.3-src.zip/py4j/java_gat...	2016/11/24 19:51:30	38 ms	2/2	<div>9/9</div>
344	runJob at PythonRDD.scala:441	2016/11/24 19:51:30	0.1 s	1/1 (1 skipped)	<div>8/8 (1 skipped)</div>
343	runJob at PythonRDD.scala:441	2016/11/24	87 ms	2/2	<div>2/2</div>

3. The screenshot of results (word, freq) printed in your console when you start putting data into Kafka.

```
Nehas-MacBook-Pro:cloud-mini-hw4 nirmala$ zkserver start
ZooKeeper JMX enabled by default
Using config: /usr/local/etc/zookeeper/zoo.cfg
Starting zookeeper ... STARTED
```

python3 producer.py

```
sent : 801929617934159873 , Can you recommend anyone for this #job in #StLouis, MO? https://t.co/fBq8xSOREz #Hiring #CareerArc
sent : 801929622602465282 , If you're looking for work in #Phoenix, AZ, check out this #job: https://t.co/xbBHmp0soe #Healthcare #Hiring #CareerArc
sent : 801929623147687936 , Interested in a #job in #Waterloo, IA? This could be a great fit: https://t.co/X2oh8usshu #Hiring #CareerArc
sent : 801929624351444992 , Can you recommend anyone for this #job in #Midland, TX? https://t.co/HEBYzhsIhg #Clerical #Hiring #CareerArc
sent : 801929625198731264 , This #job might be a great fit for you: Software Engineer, PS COE - https://t.co/e6UeCf5tab #IT #Sofia, Sofia City Province #Hiring
sent : 801929643309744128 , current weather in Pasadena: clear sky, 72°F 56% humidity, wind 7mph, pressure 1021mb
sent : 801929652012916736 , Want to work in #MineralWells, TX? View our latest opening: https://t.co/b0WNWalntR #Job #Jobs #Hiring #CareerArc
sent : 801929656970510337 , Our American EXISTENCE taught us PUMPKINS ain't JUST a HALLOWEEN thing! Celebrate the lead up to... https://t.co/wrsD4QKy5G
sent : 801929657905856516 , Want to work in #LittleRock, AR? View our latest opening: https://t.co/6cNasXr8rz #Job #Nursing #Jobs #Hiring
sent : 801929671226986496 , Can you recommend anyone for this #job in #StPetersburg, FL? https://t.co/Vbpnnesn2m #PracticeWithUs #Physician... https://t.co/N5FHqJ1RIn
sent : 801929679347122176 , current weather in Texas City: clear sky, 71°F 56% humidity, wind 7mph, pressure 1021mb
```

Below Spark consumer is set to aggregate the RDDs every 10 seconds, so we see the counts of words in filtered topics every 10 seconds

\$(Added SparkstreamConsumer.py to the Spark installation folder)

```
./bin/spark-submit --packages org.apache.spark:spark-streaming-kafka-0-8_2.11:2.0.2 SparkstreamConsumer.py
```

```

spark-2.0.2-bin-hadoop2.7 — python • java -cp ~/spark-2.0.2-bin-hadoop2.7/conf:/Users/nirmala/spark-2.0.2-bin-ha...
8, free 365.0 MB)
16/11/24 19:41:30 INFO BlockManagerInfo: Added broadcast_383_piece0 in memory on 10.40.190.210:59355 (size: 22.
9 KB, free: 366.0 MB)
16/11/24 19:41:30 INFO SparkContext: Created broadcast 383 from runJob at PythonRDD.scala:441
16/11/24 19:41:30 INFO ReliableRDDCheckpointData: Done checkpointing RDD 1197 to file:/Users/nirmala/spark-2.0.
2-bin-hadoop2.7/checkpoint/d8681a27-4fc6-4c11-9318-e09034ded5e5/rdd-1197, new parent is RDD 1204
=====
Time: 2016-11-24 19:41:30
-----
('', 3)
('Can', 32)
('801947234849853440', 1)
('Smurf', 1)
('Roslindale', 1)
('801947832697532416', 1)
('meant!', 1)
('CNN', 1)
('https://t.co/o38I43pJn0', 1)
('1480034026841', 1)
...

16/11/24 19:41:30 INFO JobScheduler: Finished job streaming job 1480034490000 ms.0 from job set of time 1480034
490000 ms
16/11/24 19:41:30 INFO JobScheduler: Starting job streaming job 1480034490000 ms.1 from job set of time 1480034
490000 ms
16/11/24 19:41:30 INFO SparkContext: Starting job: call at /Users/nirmala/spark-2.0.2-bin-hadoop2.7/python/lib/
py4j-0.10.3-src.zip/py4j/java_gateway.py:2230
16/11/24 19:41:30 INFO DAGScheduler: Registering RDD 1199 (call at /Users/nirmala/spark-2.0.2-bin-hadoop2.7/pyt
hon/lib/py4j-0.10.3-src.zip/py4j/java_gateway.py:2230)
16/11/24 19:41:30 INFO DAGScheduler: Got job 165 (call at /Users/nirmala/spark-2.0.2-bin-hadoop2.7/python/lib/p
y4j-0.10.3-src.zip/py4j/java_gateway.py:2230) with 8 output partitions
16/11/24 19:41:30 INFO DAGScheduler: Final stage: ResultStage 330 (call at /Users/nirmala/spark-2.0.2-bin-hadoo
p2.7/python/lib/py4j-0.10.3-src.zip/py4j/java_gateway.py:2230)
16/11/24 19:41:30 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 329)

```

Count of words gets updated

```

2-bin-hadoop2.7/checkpoint/d8681a27-4fc6-4c11-9318-e09034ded5e5/rdd-1241, new parent is RDD 1248
=====
Time: 2016-11-24 19:41:50
-----
('', 5)
('Can', 36)
('801947234849853440', 1)
('Smurf', 1)
('Roslindale', 1)
('801947832697532416', 1)
('meant!', 1)
('CNN', 1)
('https://t.co/o38I43pJn0', 1)
('1480034026841', 1)
...

16/11/24 19:41:50 INFO JobScheduler: Finished job streaming job 1480034510000 ms.0 from job set of time 1480034
510000 ms
16/11/24 19:41:50 INFO JobScheduler: Starting job streaming job 1480034510000 ms.1 from job set of time 1480034
510000 ms
16/11/24 19:41:50 INFO SparkContext: Starting job: call at /Users/nirmala/spark-2.0.2-bin-hadoop2.7/python/lib/
py4j-0.10.3-src.zip/py4j/java_gateway.py:2230

```