

Explanation for Spark Streaming in Lab2

Implementation details and the whole framework:

There are three parts in code.

- 1) Data source
- 2) Data processing
- 3) Data Storing

Data source: Kafka. Set configuration of Kafka and build integration with Spark Streaming. Then Use Receiver-less direct approach (by calling `KafkaUtils.createDirectStream`) to read data.

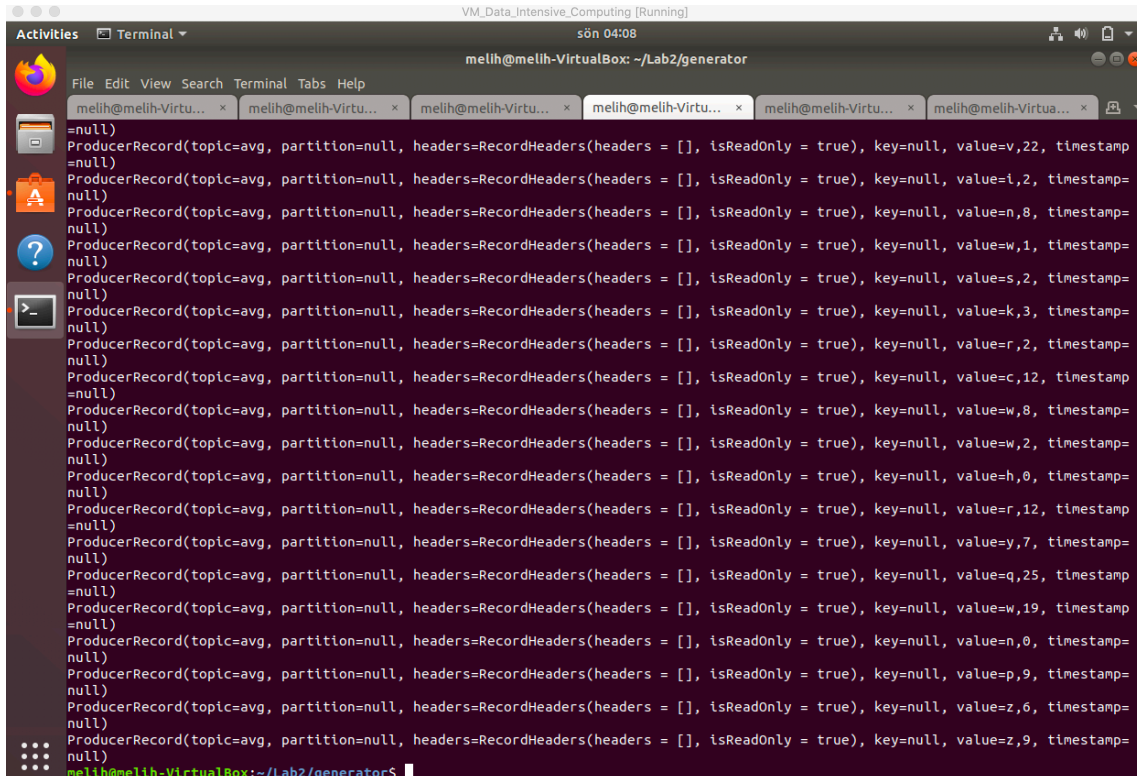
Data processing: Spark Streaming. Set configuration and create `StreamingContext`. Read data from Kafka as `Dstream`. Process it as (key, value) pairs by map function-> (String, Int). Use `mapWithState` to calculate the average. In update function (mappingFunc [key, value, state]), define state as (Double, Int). Int type stores old count, Double type stores old average. Then update state every time new data coming, and update the average value of each key by calculating $\text{newAvg} = (\text{oldAvg} * \text{oldCount} + \text{newVal}) / \text{newCount}$. (every time $\text{newCount} = \text{oldCount} + 1$ if data exists)

Data Storing: Cassandra. Build keyspace("avg_space") and create table("avg") with columns "word: text", "count: float", set connection with Spark Streaming. Finally, use `stateDstream.saveToCasandra` to store the result in Cassandra.

Run the code

Start Zookeeper -> start Kafka -> start Cassandra -> run spark streaming code by running sbt -> run generator code to produce data by running sbt -> check Cassandra table to show result

Result



```
VM_Data_Intensive_Computing [Running]
sön 04:08
melih@melih-VirtualBox: ~/Lab2/generator

File Edit View Search Terminal Tabs Help

melih@melih-Virtu... melih@melih-Virtu... melih@melih-Virtu... melih@melih-Virtu... melih@melih-Virtu... melih@melih-Virtu...

= null)
ProducerRecord(topic=avg, partition=null, headers=RecordHeaders(headers = [], isReadOnly = true), key=null, value=v,22, timestamp=
= null)
ProducerRecord(topic=avg, partition=null, headers=RecordHeaders(headers = [], isReadOnly = true), key=null, value=l,2, timestamp=
= null)
ProducerRecord(topic=avg, partition=null, headers=RecordHeaders(headers = [], isReadOnly = true), key=null, value=n,8, timestamp=
= null)
ProducerRecord(topic=avg, partition=null, headers=RecordHeaders(headers = [], isReadOnly = true), key=null, value=w,1, timestamp=
= null)
ProducerRecord(topic=avg, partition=null, headers=RecordHeaders(headers = [], isReadOnly = true), key=null, value=s,2, timestamp=
= null)
ProducerRecord(topic=avg, partition=null, headers=RecordHeaders(headers = [], isReadOnly = true), key=null, value=k,3, timestamp=
= null)
ProducerRecord(topic=avg, partition=null, headers=RecordHeaders(headers = [], isReadOnly = true), key=null, value=r,2, timestamp=
= null)
ProducerRecord(topic=avg, partition=null, headers=RecordHeaders(headers = [], isReadOnly = true), key=null, value=c,12, timestamp=
= null)
ProducerRecord(topic=avg, partition=null, headers=RecordHeaders(headers = [], isReadOnly = true), key=null, value=w,8, timestamp=
= null)
ProducerRecord(topic=avg, partition=null, headers=RecordHeaders(headers = [], isReadOnly = true), key=null, value=w,2, timestamp=
= null)
ProducerRecord(topic=avg, partition=null, headers=RecordHeaders(headers = [], isReadOnly = true), key=null, value=h,0, timestamp=
= null)
ProducerRecord(topic=avg, partition=null, headers=RecordHeaders(headers = [], isReadOnly = true), key=null, value=r,12, timestamp=
= null)
ProducerRecord(topic=avg, partition=null, headers=RecordHeaders(headers = [], isReadOnly = true), key=null, value=y,7, timestamp=
= null)
ProducerRecord(topic=avg, partition=null, headers=RecordHeaders(headers = [], isReadOnly = true), key=null, value=q,25, timestamp=
= null)
ProducerRecord(topic=avg, partition=null, headers=RecordHeaders(headers = [], isReadOnly = true), key=null, value=w,19, timestamp=
= null)
ProducerRecord(topic=avg, partition=null, headers=RecordHeaders(headers = [], isReadOnly = true), key=null, value=n,0, timestamp=
= null)
ProducerRecord(topic=avg, partition=null, headers=RecordHeaders(headers = [], isReadOnly = true), key=null, value=p,9, timestamp=
= null)
ProducerRecord(topic=avg, partition=null, headers=RecordHeaders(headers = [], isReadOnly = true), key=null, value=z,6, timestamp=
= null)
ProducerRecord(topic=avg, partition=null, headers=RecordHeaders(headers = [], isReadOnly = true), key=null, value=z,9, timestamp=
= null)
melih@melih-VirtualBox:~/Lab2/generator$
```

Generating data

```
VM_Data_Intensive_Computing [Running]
sön 04:09
melih@melih-VirtualBox: ~/Lab2/sparkstreaming

20/10/11 04:06:41 INFO DAGScheduler: Got job 21 (runJob at DStreamFunctions.scala:54) with 2 output partitions
20/10/11 04:06:41 INFO DAGScheduler: Final stage: ResultStage 173 (runJob at DStreamFunctions.scala:54)
20/10/11 04:06:41 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 171, ShuffleMapStage 168, ShuffleMapStage 165, ShuffleMapStage 172, ShuffleMapStage 169, ShuffleMapStage 166, ShuffleMapStage 170, ShuffleMapStage 163, ShuffleMapStage 167, ShuffleMapStage 164)
20/10/11 04:06:41 INFO DAGScheduler: Missing parents: List()
20/10/11 04:06:41 INFO DAGScheduler: Submitting ResultStage 173 (MapWithStateRDD[105] at mapWithState at KafkaSpark.scala:80), which has no missing parents
20/10/11 04:06:41 INFO MemoryStore: Block broadcast_44 stored as values in memory (estimated size 8.9 KB, free 407.9 MB)
20/10/11 04:06:41 INFO MemoryStore: Block broadcast_44_piece0 stored as bytes in memory (estimated size 4.1 KB, free 407.9 MB)
20/10/11 04:06:41 INFO BlockManagerInfo: Added broadcast_44_piece0 in memory on 10.0.2.15:42483 (size: 4.1 KB, free: 408.4 MB)
20/10/11 04:06:41 INFO SparkContext: Created broadcast 44 from broadcast at DAGScheduler.scala:1006
20/10/11 04:06:41 INFO DAGScheduler: Submitting 2 missing tasks from ResultStage 173 (MapWithStateRDD[105] at mapWithState at KafkaSpark.scala:80) (first 15 tasks are for partitions Vector(0, 1))
20/10/11 04:06:41 INFO TaskSchedulerImpl: Adding task set 173.0 with 2 tasks
20/10/11 04:06:41 INFO TaskSetManager: Starting task 0.0 in stage 173.0 (TID 62, localhost, executor driver, partition 0, PROCESS_LOCAL, 5073 bytes)
20/10/11 04:06:41 INFO TaskSetManager: Starting task 1.0 in stage 173.0 (TID 63, localhost, executor driver, partition 1, PROCESS_LOCAL, 5073 bytes)
20/10/11 04:06:41 INFO Executor: Running task 0.0 in stage 173.0 (TID 62)
20/10/11 04:06:41 INFO Executor: Running task 1.0 in stage 173.0 (TID 63)
20/10/11 04:06:41 INFO BlockManager: Found block rdd_105_1 locally
20/10/11 04:06:41 INFO BlockManager: Found block rdd_105_0 locally
20/10/11 04:06:41 INFO Executor: Finished task 0.0 in stage 173.0 (TID 62). 794 bytes result sent to driver
20/10/11 04:06:41 INFO TaskSetManager: Finished task 0.0 in stage 173.0 (TID 62) in 135 ms on localhost (executor driver) (1/2)
20/10/11 04:06:41 INFO Executor: Finished task 1.0 in stage 173.0 (TID 63). 794 bytes result sent to driver
20/10/11 04:06:41 INFO TaskSetManager: Finished task 1.0 in stage 173.0 (TID 63) in 140 ms on localhost (executor driver) (2/2)
20/10/11 04:06:41 INFO TaskSchedulerImpl: Removed TaskSet 173.0, whose tasks have all completed, from pool
20/10/11 04:06:41 INFO DAGScheduler: ResultStage 173 (runJob at DStreamFunctions.scala:54) finished in 0.119 s
20/10/11 04:06:41 INFO DAGScheduler: Job 21 finished: runJob at DStreamFunctions.scala:54, took 0.219575 s
20/10/11 04:06:41 INFO ReliableCheckpointRDD: Checkpointing took 320 ms.
20/10/11 04:06:41 INFO MemoryStore: Block broadcast_45 stored as values in memory (estimated size 214.5 KB, free 407.7 MB)
20/10/11 04:06:41 INFO MemoryStore: Block broadcast_45_piece0 stored as bytes in memory (estimated size 20.4 KB, free 407.6 MB)
20/10/11 04:06:41 INFO BlockManagerInfo: Added broadcast_45_piece0 in memory on 10.0.2.15:42483 (size: 20.4 KB, free: 408.4 MB)
20/10/11 04:06:41 INFO SparkContext: Created broadcast 45 from runJob at DStreamFunctions.scala:54
20/10/11 04:06:41 INFO ReliableRDDCheckpointData: Done checkpointing RDD 105 to file:/home/melih/Lab2/sparkstreaming/13c0117b-e30b-4841-8211-ce9a3937d83/rdd-105, new parent is RDD 107
20/10/11 04:06:41 INFO JobScheduler: Finished job streaming job 1602382000000 ms.0 from job set of time 1602382000000 ms
20/10/11 04:06:41 INFO JobScheduler: Total delay: 1.302 s for time 1602382000000 ms (execution: 1.254 s)
20/10/11 04:06:41 INFO MapPartitionsRDD: Removing RDD 101 from persistence list
```

Spark Streaming running process

```
Activities Terminal
sön 04:07
melih@melih-VirtualBox: ~

InvalidRequest: Error from server: code=2200 [Invalid query] message="unconfigured table words"
cqlsh:avg_space> select * from avg;

word | count
-----+-----
(0 rows)
cqlsh:avg_space> select * from avg;

word | count
-----+-----
z | 12.15576
a | 12.64878
c | 12.70819
m | 12.18524
f | 12.2178
o | 12.59241
n | 12.21842
q | 12.64715
g | 12.38757
p | 12.8437
e | 12.59315
r | 12.59351
d | 12.2529
h | 12.65155
w | 12.11308
l | 12.34372
j | 12.27722
v | 12.39201
y | 12.4615
u | 12.61679
i | 12.65485
k | 12.96581
t | 12.3395
x | 12.38762
b | 12.37151
s | 12.40495

(26 rows)
cqlsh:avg_space>
```

Final result in Cassandra