Coding Assignment

Table of Contents

For below mentioned exercise, share relevant code and snapshots

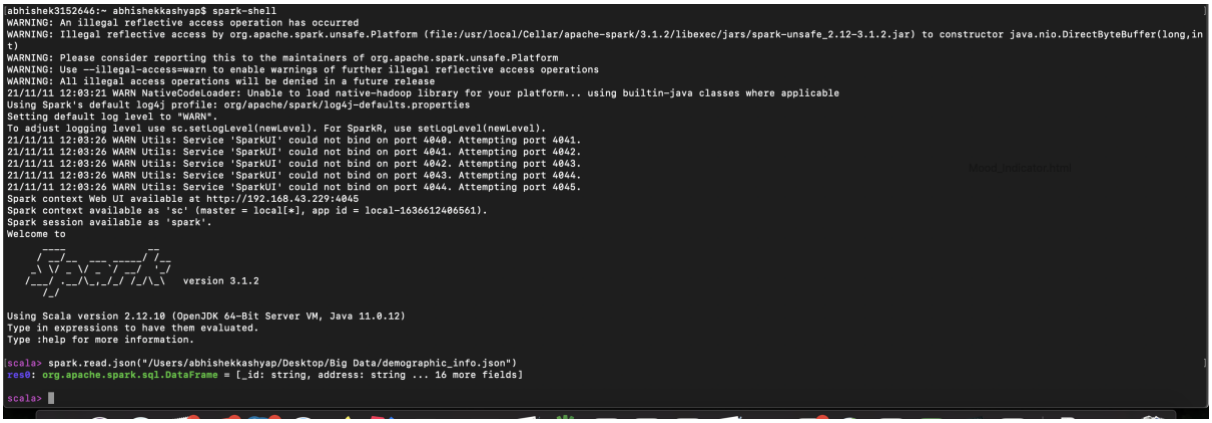
Spark Batch

1. Read attached json (demographic_info.json) into a Spark Dataframe and carry out following –

Query:

```
spark.read.json("/Users/abhishekkashyap/Desktop/Big Data/demographic_info.json")
```

Screenshot:



2. Use following apis over Dataframe **select** to show columns (name, age, gender, isActive, balance, company, eyeColor, email, phone) **filter** to show records with isActive as true

Query:

```
res0.filter("isActive == 'true'").select("name", "age", "gender", "isActive", "balance", "company", "eyeColor", "email", "phone").show()
```

Screenshot:

3. Show top 2 male and female with maximum balance

Query:

```
import org.apache.spark.sql.expressions.Window

val win=Window.partitionBy("gender").orderBy(desc("balance"))

res0.withColumn("rank",
rank().over(win).alias("rank")).filter(col("rank") <= 2).show()
```

Screenshot:

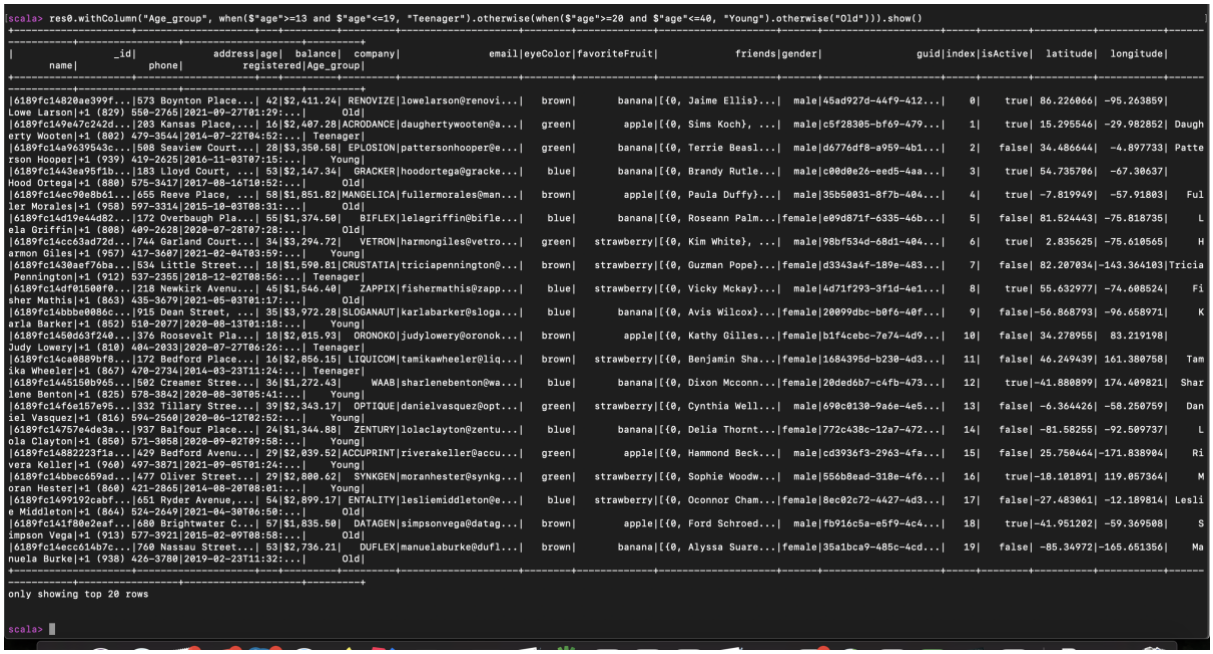
4. Add a column Age_group with classifications as Teenager (13-19 years), Young (20-40 years), Old (>40 years)

Query:

```
res0.withColumn("Age_group", when($"age">=13 and $"age"<=19,
"Teenager").otherwise(when($"age">=20 and $"age"<=40,
"Young").otherwise("Old"))).show()
```

Screenshot:


```
res0.withColumn("Age_group", when($"age">=13 and $"age"<=19,
"Teenager").otherwise(when($"age">=20 and $"age"<=40,
"Young").otherwise("Old"))).show()
```

5. Create temp table view over initially read json Dataframe and run sql queries for same requirements given above.

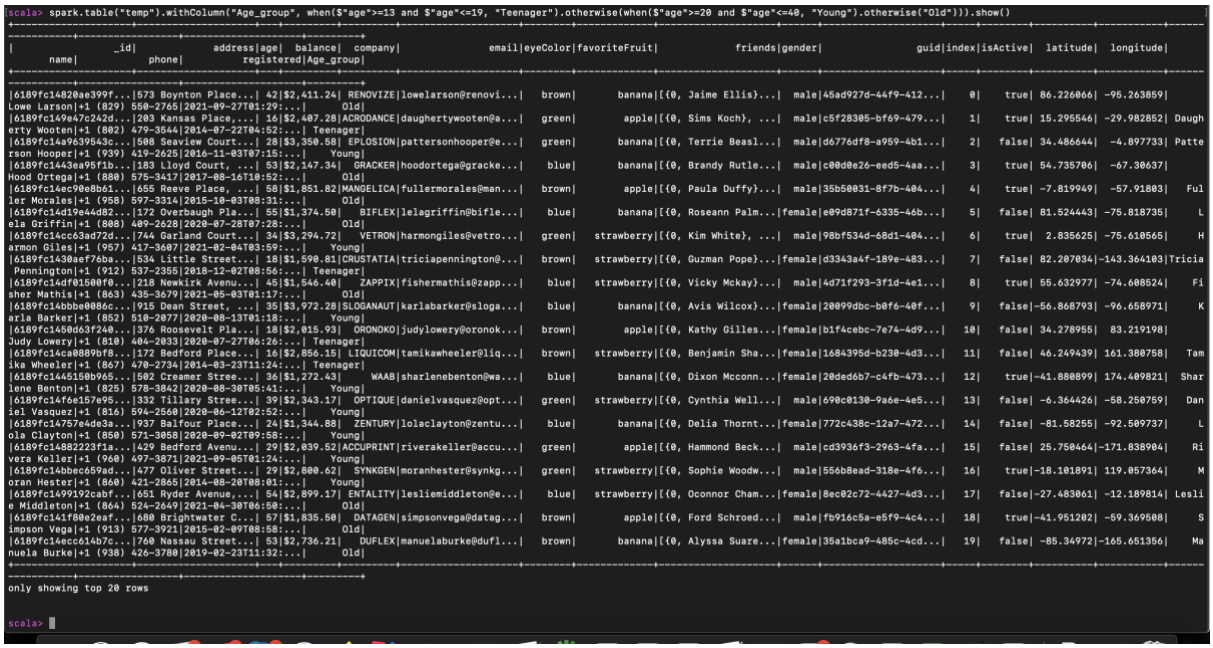
Query:

```
res0.createOrReplaceTempView("temp")

spark.table("temp").filter("isActive == 'true'").select("name",
"age", "gender", "isActive", "balance", "company", "eyeColor",
"email", "phone").show()
```

Screenshot


```
val win=Window.partitionBy("gender").orderBy(desc("balance"))

spark.table("temp").withColumn("rank",
rank().over(win).alias("rank")).filter(col("rank") <= 2).show()
```

```
spark.table("temp").withColumn("Age_group", when($"age">=13 and $"age"<=19, "Teenager").otherwise(when($"age">=20 and $"age"<=40, "Young").otherwise("Old"))).show()
```



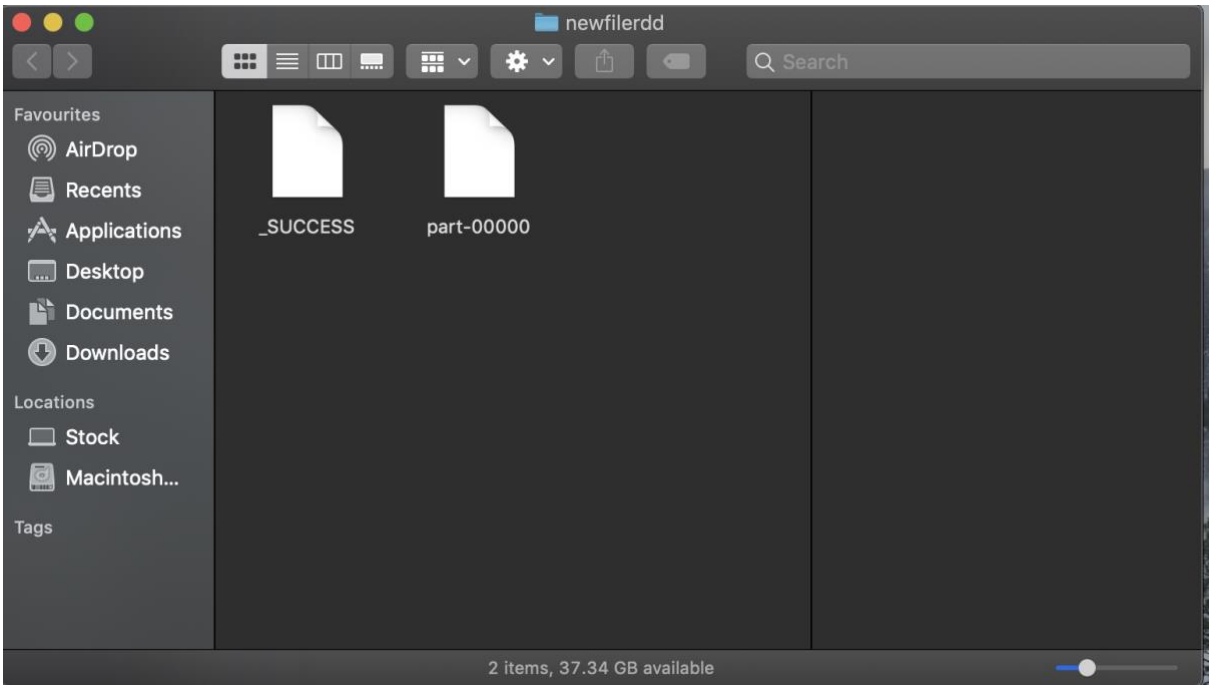
```
scala> spark.table("temp").withColumn("Age_group", when($"age">=13 and $"age"<=19, "Teenager").otherwise(when($"age">=20 and $"age"<=40, "Young").otherwise("Old"))).show()
+--------------------+--------------------+--------------------+---+--------+---------+--------------------+--------+-------------+--------------------+------+--------------------+-----+--------+---------+-----------+------+
|                name|               phone|                 _id|            address|age| balance|  company|               email|eyeColor|favoriteFruit|             friends|gender|                guid|index|isActive| latitude|  longitude|Age_group|
+--------------------+--------------------+--------------------+-------------------+---+--------+---------+--------------------+--------+-------------+--------------------+------+--------------------+-----+--------+---------+-----------+------+
|Lowe Larson|+1 (829) 550-2765|6189fc14820ae399f...|673 Boynton Place...| 42|$2,411.24| RENOVIZE|lowelarson@renovi...|   brown|       banana|[{0, Jaime Ellis}...|  male|45ad927d-44f9-412...|    0|    true| 86.226066| -95.263859|      Old|
|erty Wooten|+1 (802) 479-3544|6189fc149e47c242d...|203 Kansas Place,...| 16|$2,407.28|ACRODANCE|daughertywooten@a...|   green|        apple|[{0, Sims Koch}, ...|  male|c5f28305-bf69-479...|    1|    true| 15.295546| -29.982852|  Teenager|
|rson Hooper|+1 (939) 419-2625|6189fc14a9639543c...|508 Seaview Court...| 28|$3,350.58| EPLOSION|pattersonhooper@e...|   green|       banana|[{0, Terrie Beasl...|  male|d6776df8-a959-4b1...|    2|   false| 34.486644|  -4.897733|  Young|
|Hood Ortega|+1 (880) 575-3417|6189fc1443ea95f1b...|183 Lloyd Court, ...| 53|$2,147.34|  GRACKER|hoodortega@gracke...|    blue|       banana|[{0, Brandy Rutle...|  male|c00d0e26-eed5-4aa...|    3|    true| 54.735706|  -67.30637|      Old|
|ler Morales|+1 (958) 597-3314|6189fc14ec90e8b61...|655 Reeve Place, ...| 58|$1,851.82|MANGELICA|fullermorales@man...|   brown|        apple|[{0, Paula Duffy}...|  male|35b50031-8f7b-404...|    4|    true| -7.819949|  -57.91803|      Old|
|ela Griffin|+1 (808) 409-2628|6189fc14d19e44d82...|172 Overbaugh Pla...| 55|$1,374.50|   BIFLEX|lelagriffin@bifle...|    blue|       banana|[{0, Roseann Palm...|female|e09d871f-6335-46b...|    5|   false| 81.524443| -75.818735|  Young|
|armon Giles|+1 (957) 417-3607|6189fc14cc63ad72d...|744 Garland Court...| 34|$3,294.72|   VETRON|harmongiles@vetro...|   green|  strawberry|[{0, Kim White}, ...|  male|98bf534d-68d1-404...|    6|    true|  2.835625| -75.610565|      Old|
|  Pennington|+1 (912) 537-2355|6189fc1430aef76ba...|534 Little Street...| 18|$1,590.81|CRUSTATIA|triciapennington@...|   brown|  strawberry|[{0, Guzman Pope}...|female|d3343a4f-189e-483...|    7|   false| 82.207034|-143.364103|  Teenager|
|sher Mathis|+1 (863) 435-3679|6189fc14df01500f0...|218 Newkirk Avenu...| 45|$1,546.40|   ZAPPIX|fishermathis@zapp...|    blue|  strawberry|[{0, Vicky Mckay}...|  male|4d71f293-3f1d-4e1...|    8|    true| 55.632977| -74.608524|  Young|
|arla Barker|+1 (852) 510-2077|6189fc1450d63f240...|376 Roosevelt Pla...| 18|$2,015.93|  ORONOKO|judylowery@oronok...|   brown|        apple|[{0, Kathy Gilles...|female|b1f4cebc-7e74-4d9...|   10|   false| 34.278955|  83.219198|      Old|
|Judy Lowery|+1 (810) 404-2033|6189fc14ca0889bf8...|172 Bedford Place...| 16|$2,856.15| LIQUICOM|tamikawheeler@liq...|   brown|  strawberry|[{0, Benjamin Sha...|female|1684395d-b230-4d3...|   11|   false| 46.249439| 161.380758|  Teenager|
|ika Wheeler|+1 (867) 470-2734|6189fc1445150b965...|502 Creamer Stree...| 36|$1,272.43|     WAAB|sharlenebenton@wa...|    blue|       banana|[{0, Dixon Mcconn...|female|20ded6b7-c4fb-473...|   12|    true|-41.880899| 174.409821|  Young|
|iel Vasquez|+1 (816) 594-2560|6189fc14f6e157e95...|332 Tillary Stree...| 39|$2,343.17|  OPTIQUE|danielvasquez@opt...|   green|  strawberry|[{0, Cynthia Well...|  male|690c0130-9a6e-4e5...|   13|   false| -6.364426| -58.250759|  Young|
|ola Clayton|+1 (850) 571-3058|6189fc14757e4de3a...|937 Balfour Place...| 24|$1,344.88|  ZENTURY|lolaclayton@zentu...|    blue|       banana|[{0, Delia Thornt...|female|772c438c-12a7-472...|   14|   false| -81.58255| -92.589737|      Old|
|vera Keller|+1 (960) 497-3871|6189fc14882223f1a...|429 Bedford Avenu...| 29|$2,039.52|ACCUPRINT|riverakeller@accu...|   green|        apple|[{0, Hammond Beck...|  male|cd3936f3-2963-4fa...|   15|   false| 25.750464|-171.838904|  Young|
|oran Hester|+1 (860) 421-2865|6189fc14bbec659ad...|477 Oliver Street...| 29|$2,800.62|  SYNKGEN|moranhester@synkg...|   green|  strawberry|[{0, Sophie Woodw...|  male|556b8ead-318e-4f6...|   16|    true|-18.101891| 119.057364|  Young|
|e Middleton|+1 (864) 524-2649|6189fc1499192cabf...|651 Ryder Avenue,...| 54|$2,899.17| ENTALITY|lesliemiddleton@e...|    blue|  strawberry|[{0, Oconnor Cham...|female|8ec02c72-4427-4d3...|   17|   false|-27.483061| -12.189814|  Young|
|impson Vega|+1 (913) 577-3921|6189fc141f80e2eaf...|680 Brightwater C...| 57|$1,835.50|  DATAGEN|simpsonvega@datag...|   brown|        apple|[{0, Ford Schroed...|  male|fb916c5a-e5f9-4c4...|   18|    true|-41.951202| -59.369508|      Old|
|nuela Burke|+1 (938) 426-3780|6189fc14ecc614b7c...|760 Nassau Street...| 53|$2,736.21|   DUFLEX|manuelaburke@dufl...|   brown|       banana|[{0, Alyssa Suare...|female|35a1bca9-485c-4cd...|   19|   false| -85.34972|-165.651356|      Old|
+--------------------+--------------------+--------------------+-------------------+---+--------+---------+--------------------+--------+-------------+--------------------+------+--------------------+-----+--------+---------+-----------+------+
only showing top 20 rows

scala>
```

6. Convert above selected column dataframe into an RDD and save it into a text file.

Query:

```
res9.rdd.saveAsTextFile("/Users/abhishekkashyap/Desktop/newfilerdd")
```

Screenshot:



Kafka

Start Server

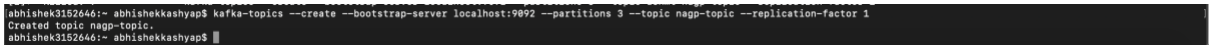
```
zookeeper-server-start /usr/local/etc/kafka/zookeeper.properties &
kafka-server-start /usr/local/etc/kafka/server.properties
```

1. Create a topic with 3 partitions

Query:

```
kafka-topics --create --bootstrap-server localhost:9092 --partitions
3 --topic nagp-topic --replication-factor 1
```

Screenshot:

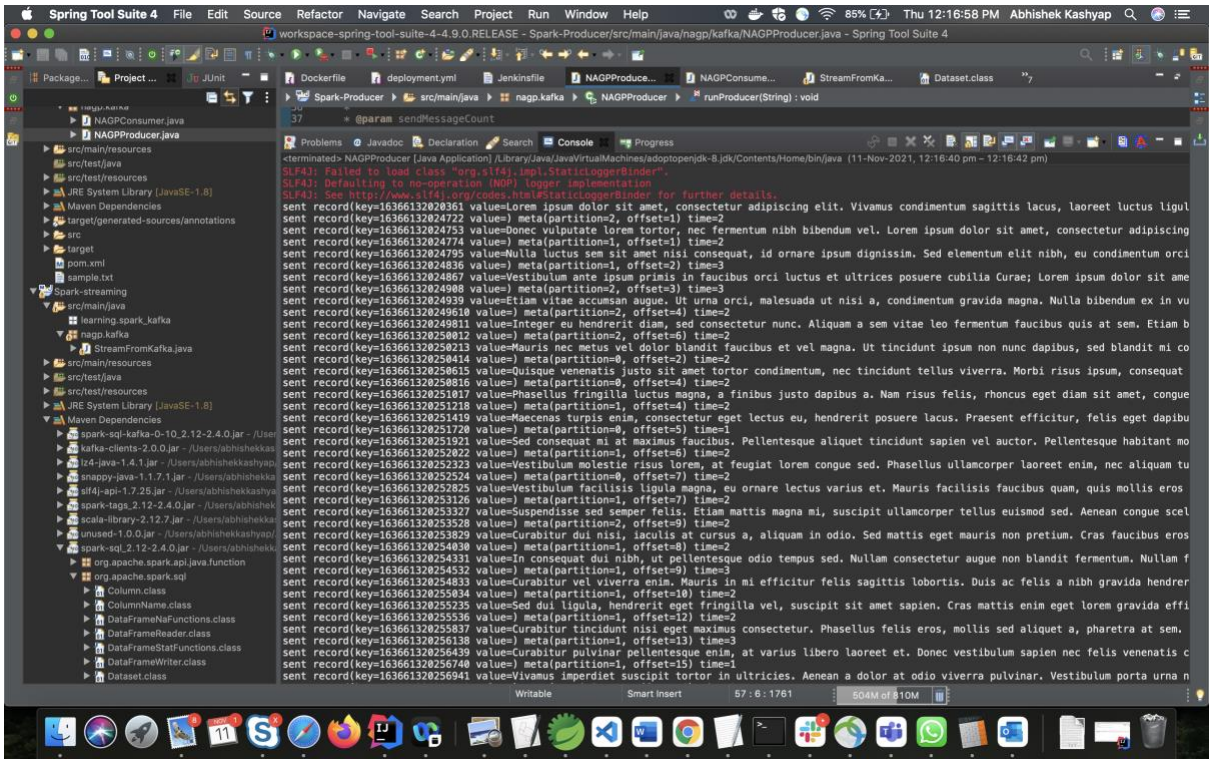


2. Create a producer writing data to above created kafka topic with following considerations –
   - 2.1 messages should be read from a file line by line (use any file from your side with limited content)
   - 2.2 message should be produced on kafka topic in (key, Value) format where key is timestamp+index and value is actual message

Code:

NAGPProducer java file in attached Spark-Producer project.

Screenshot:


Spring Tool Suite 4 File Edit Source Refactor Navigate Search Project Run Window Help
workspace-spring-tool-suite-4-4.9.0.RELEASE - Spark-Producer/src/main/java/nagp/kafka/NAGPProducer.java - Spring Tool Suite 4
Spark-Producer ▸ src/main/java ▸ nagp.kafka ▸ NAGPProducer ▸ runProducer(String) : void
@param sendMessageCount
Problems Javadoc Declaration Search Console Progress
<terminated> NAGPProducer [Java Application] /Library/Java/JavaVirtualMachines/adoptopenjdk-8.jdk/Contents/Home/bin/java (11-Nov-2021, 12:16:40 pm – 12:16:42 pm)
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
sent record(key=163661320203 61) value=Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vivamus condimentum sagittis lacus, laoreet luctus ligul
sent record(key=163661320247 22 value=) meta(partition=2, offset=1) time=2
sent record(key=163661320247 53 value=Donec vulputate lorem tortor, nec fermentum nibh bibendum vel. Lorem ipsum dolor sit amet, consectetur adipiscing
sent record(key=163661320247 74 value=) meta(partition=1, offset=1) time=2
sent record(key=163661320247 95 value=Nulla luctus sem sit amet nisi consequat, id ornare ipsum dignissim. Sed elementum elit nibh, eu condimentum orci
sent record(key=163661320248 36 value=) meta(partition=1, offset=2) time=3
sent record(key=163661320248 67 value=Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Lorem ipsum dolor sit ame
sent record(key=163661320249 08 value=) meta(partition=2, offset=3) time=3
sent record(key=163661320249 39 value=Etiam vitae accumsan augue. Ut urna orci, malesuada ut nisi a, condimentum gravida magna. Nulla bibendum ex in vu
sent record(key=163661320249 610 value=) meta(partition=2, offset=4) time=2
sent record(key=163661320249 811 value=Integer eu hendrerit diam, sed consectetur nunc. Aliquam a sem vitae leo fermentum faucibus quis at sem. Etiam b
sent record(key=163661320250 012 value=) meta(partition=2, offset=6) time=2
sent record(key=163661320250 213 value=Mauris nec metus vel dolor blandit faucibus et vel magna. Ut tincidunt ipsum non nunc dapibus, sed blandit mi co
sent record(key=163661320250 414 value=) meta(partition=0, offset=2) time=2
sent record(key=163661320250 615 value=Quisque venenatis justo sit amet tortor condimentum, nec tincidunt tellus viverra. Morbi risus ipsum, consequat
sent record(key=163661320250 816 value=) meta(partition=0, offset=4) time=2
sent record(key=163661320251 017 value=Phasellus fringilla luctus magna, a finibus justo dapibus a. Nam risus felis, rhoncus eget diam sit amet, congue
sent record(key=163661320251 218 value=) meta(partition=1, offset=4) time=2
sent record(key=163661320251 419 value=Maecenas turpis enim, consectetur eget lectus eu, hendrerit posuere lacus. Praesent efficitur, felis eget dapibu
sent record(key=163661320251 720 value=) meta(partition=0, offset=5) time=1
sent record(key=163661320251 921 value=Sed consequat mi at maximus faucibus. Pellentesque aliquet tincidunt sapien vel auctor. Pellentesque habitant mo
sent record(key=163661320252 022 value=) meta(partition=1, offset=6) time=2
sent record(key=163661320252 323 value=Vestibulum molestie risus lorem, at feugiat lorem congue sed. Phasellus ullamcorper laoreet enim, nec aliquam tu
sent record(key=163661320252 524 value=) meta(partition=0, offset=7) time=2
sent record(key=163661320252 825 value=Vestibulum facilisis ligula magna, eu ornare lectus varius et. Mauris facilisis faucibus quam, quis mollis eros
sent record(key=163661320253 126 value=) meta(partition=1, offset=7) time=2
sent record(key=163661320253 327 value=Suspendisse sed semper felis. Etiam mattis magna mi, suscipit ullamcorper tellus euismod sed. Aenean congue scel
sent record(key=163661320253 528 value=) meta(partition=2, offset=9) time=2
sent record(key=163661320253 829 value=Curabitur dui nisi, iaculis at cursus a, aliquam in odio. Sed mattis eget mauris non pretium. Cras faucibus eros
sent record(key=163661320254 030 value=) meta(partition=1, offset=8) time=2
sent record(key=163661320254 331 value=In consequat dui nibh, ut pellentesque odio tempus sed. Nullam consectetur augue non blandit fermentum. Nullam f
sent record(key=163661320254 532 value=) meta(partition=1, offset=9) time=3
sent record(key=163661320254 833 value=Curabitur vel viverra enim. Mauris in mi efficitur felis sagittis lobortis. Duis ac felis a nibh gravida hendrer
sent record(key=163661320255 034 value=) meta(partition=1, offset=10) time=2
sent record(key=163661320255 235 value=Sed dui ligula, hendrerit eget fringilla vel, suscipit sit amet sapien. Cras mattis enim eget lorem gravida effi
sent record(key=163661320255 536 value=) meta(partition=1, offset=12) time=2
sent record(key=163661320255 837 value=Curabitur tincidunt nisi eget maximus consectetur. Phasellus felis eros, mollis sed aliquet a, pharetra at sem.
sent record(key=163661320256 138 value=) meta(partition=1, offset=13) time=3
sent record(key=163661320256 439 value=Curabitur pulvinar pellentesque enim, at varius libero laoreet et. Donec vestibulum sapien nec felis venenatis c
sent record(key=163661320256 740 value=) meta(partition=1, offset=15) time=1
sent record(key=163661320256 941 value=Vivamus imperdiet suscipit tortor in ultricies. Aenean a dolor at odio viverra pulvinar. Vestibulum porta urna n
Writable Smart Insert 57 : 6 : 1761 504M of 810M

Structured Streaming

1. Create a Spark streaming job reading data from above created kafka topic with following considerations –

1. read from beginning
2. calculate word count
3. print word count to console

Code:

StreamFromKafka java file in attached Spark-streaming project.

Screenshot: