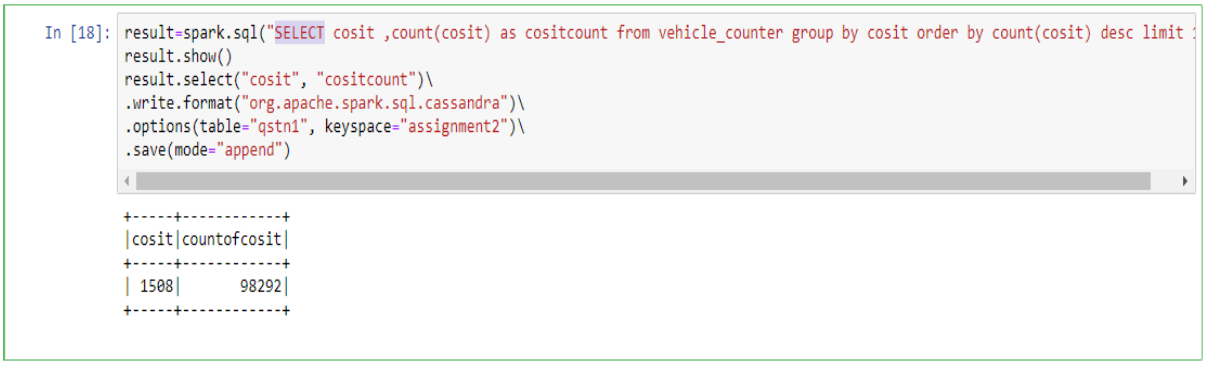
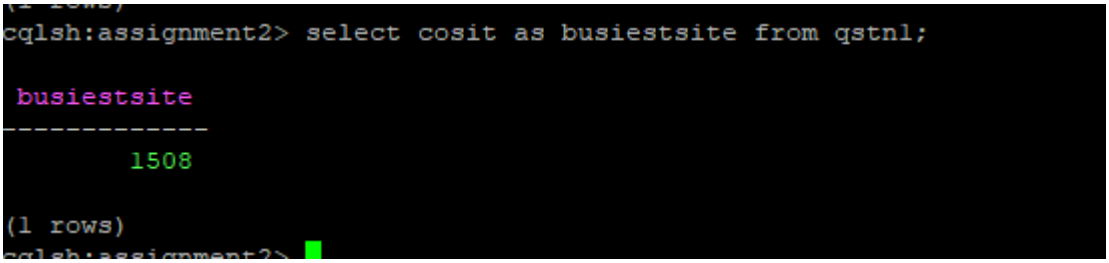
Big Data Management
Assignment 2

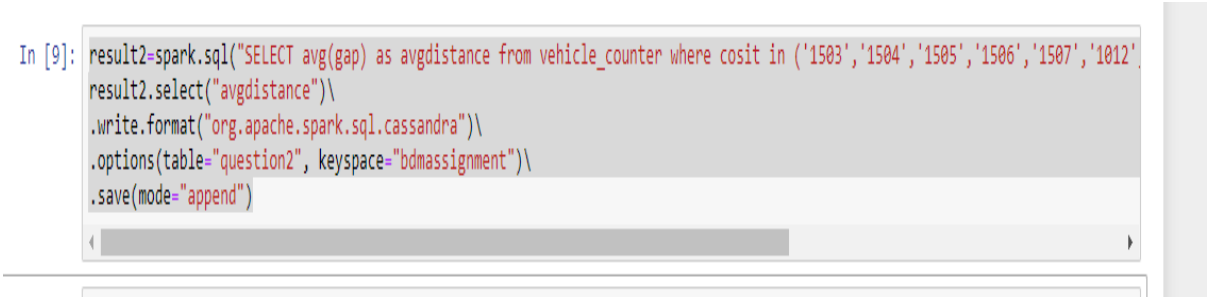
1)Show the busiest site in Ireland.

result=spark.sql("SELECT cosit ,count(cosit) as cositcount from vehicle_counter group by cosit order by count(cosit) desc limit 1")
result.show()
result.select("cosit", "cositcount")\
.write.format("org.apache.spark.sql.cassandra")\
.options(table="qstn1", keyspace="assignment2")\
.save(mode="append")





2)Show the average distance between vehicles on all M50 sites.



```
cqlsh:bdmassignment> select * from question2;

 avgdistance
--------------
      4.27486

(1 rows)
cqlsh:bdmassignment>
```

3. What site has recorded the highest temperature? Show the hour of the day.

```
result3=spark.sql("SELECT temperature, hour, cosit,day from vehicle_counter order by temperature desc limit 1")
result3.show()
result3=spark.sql("SELECT temperature, hour, cosit , day from vehicle_counter order by temperature desc limit 1")
result3.select("temperature", "hour","cosit","day")\
.write.format("org.apache.spark.sql.cassandra")\
.options(table="question3", keyspace="bdmassignment")\
.save(mode="append")
```

```
+-----------+----+-----+---+
|temperature|hour|cosit|day|
+-----------+----+-----+---+
|       12.0|  18| 1015| 31|
+-----------+----+-----+---+
```

```
cqlsh:bdmassignment> select * from question3;

 cosit | day | hour | temperature
-------+-----+------+-------------
  1015 |  31 |   18 |          12

(1 rows)
cqlsh:bdmassignment>
```

4. Show total number of WIM sites available in the dataset?

```
In [4]: result4=spark.sql("SELECT  count(cosit) as wimsitecount from vehicle_counter  where weight!=0 ")
        result4.show()
        result4=spark.sql("SELECT  count(cosit) as wimsitecount from vehicle_counter  where weight!=0 ")
        result4.select("wimsitecount")\
        .write.format("org.apache.spark.sql.cassandra")\
        .options(table="question4", keyspace="bdmassignment")\
        .save(mode="append")
```

```
+------------+
|wimsitecount|
+------------+
|      188176|
+------------+
```

```
cqlsh:bdmassignment> select * from question4;

 wimsitecount
--------------
       188176

(1 rows)
cqlsh:bdmassignment>
```

5. Compute the average speed for each site on M50.

```
]: result5=spark.sql("SELECT avg(speed) as avgspeed from vehicle_counter where cosit in ('1503','1504','1505','1506','1507','1012',
   result5.show()
   result5=spark.sql("SELECT avg(speed) as avgspeed from vehicle_counter where cosit in ('1503','1504','1505','1506','1507','1012',
   result5.select("cosit", "avgspeed ")\
   .write.format("org.apache.spark.sql.cassandra")\
   .options(table="question5", keyspace="bdmassignment")\
   .save(mode="append")
```

```
+-----------------+
|         avgspeed|
+-----------------+
|82.66201226494377|
+-----------------+
```

```
In [5]: result5=spark.sql("SELECT avg(speed) as avgspeed , cosit from vehicle_counter where cosit in ('1503','1504','1505','1506','1507'
        result5.show()
```

```
+-----------------+-----+
|         avgspeed|cosit|
+-----------------+-----+
|95.00087226856925| 1507|
+-----------------+-----+
```

6. Show total number of counts by vehicle class. Order results in descending.

```
In [7]: result6=spark.sql("SELECT  count(class) as vehiclecount , class  from vehicle_counter  group by class order by count(class) desc
        result6.show()
        result6=spark.sql("SELECT  count(class) as vehiclecount , class  from vehicle_counter  group by class order by count(class) desc
        result6.select("vehiclecount ", "class")\
        .write.format("org.apache.spark.sql.cassandra")\
        .options(table="question6", keyspace="bdmassignment")\
        .save(mode="append")
```

```
+------------+-----+
|vehiclecount|class|
+------------+-----+
|     3472965|    2|
|      498505|    3|
|      216978|    6|
|      135202|    5|
|       29347|    4|
|       21224|    7|
|       14682|    1|
|         396|    0|
+------------+-----+
```

7. List the top 3 busiest sites on M50.

```
In [8]: result7=spark.sql("SELECT cosit ,count(cosit) as cositcount from vehicle_counter where cosit in ('1014','20021','1113','1012','1
        result7.show()
        result7.select("cosit ", "cositcount")\
        .write.format("org.apache.spark.sql.cassandra")\
        .options(table="question7", keyspace="bdmassignment")\
        .save(mode="append")
```

```
+-----+----------+
|cosit|cositcount|
+-----+----------+
| 1508|     98292|
| 1502|     89498|
| 1503|     86195|
+-----+----------+
```

8. What is the busiest site on M6?

```
In [9]: result8=spark.sql("SELECT cosit ,count(cosit) as cositcount from vehicle_counter  where cosit in ('3172','3182') group by cosit
        result8.show()
        result8.select("cositcount","cosit ")\
        .write.format("org.apache.spark.sql.cassandra")\
        .options(table="question8", keyspace="bdmassignment")\
        .save(mode="append")
```

```
+-----+----------+
|cosit|cositcount|
+-----+----------+
| 3172|      9112|
+-----+----------+
```

9. What site reports the highest number of HGVs?

```
: result9=spark.sql("SELECT  count(cosit) as HGVcount, cosit  from vehicle_counter  where classname in ('HGV_ART','HGV_RIG') group
  result9.show()
  result9.select("HGVcount","cosit ","classname")\
  .write.format("org.apache.spark.sql.cassandra")\
  .options(table="question9", keyspace="bdmassignment")\
  .save(mode="append")
```

```
+--------+-----+
|HGVcount|cosit|
+--------+-----+
|   12031|  997|
+--------+-----+
```

10. Calculate the total number of vehicles on each site on M7.

```
In [2]: result10=spark.sql("SELECT  count(cosit) as numberofvehicles , cosit  from vehicle_counter  where cosit in ('20073', '20081', '2
        result10.show()
        result10.select("numberofvehicles","cosit ")\
        .write.format("org.apache.spark.sql.cassandra")\
        .options(table="question10", keyspace="bdmassignment")\
        .save(mode="append")
```

```
+---------------+------+
|numberofvehicles| cosit|
+---------------+------+
|           5521| 20081|
|          17568|200722|
+---------------+------+
```