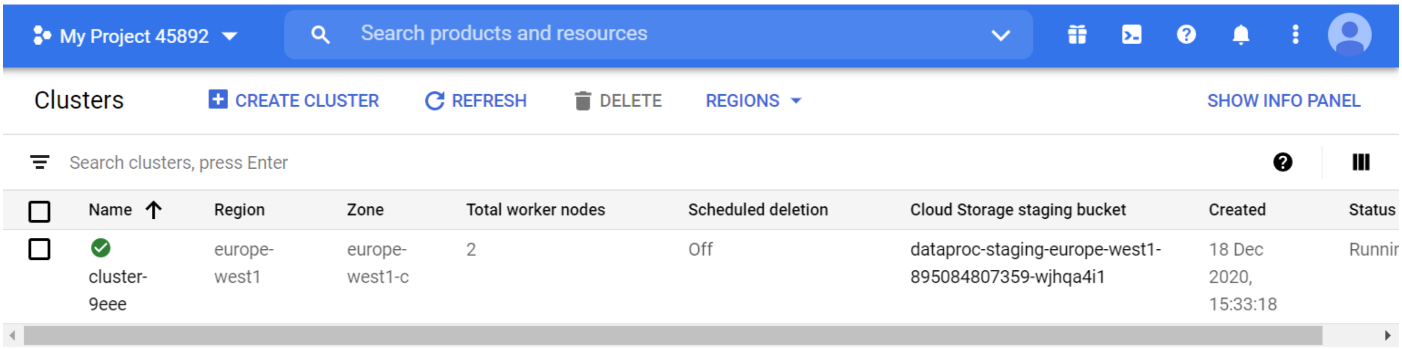
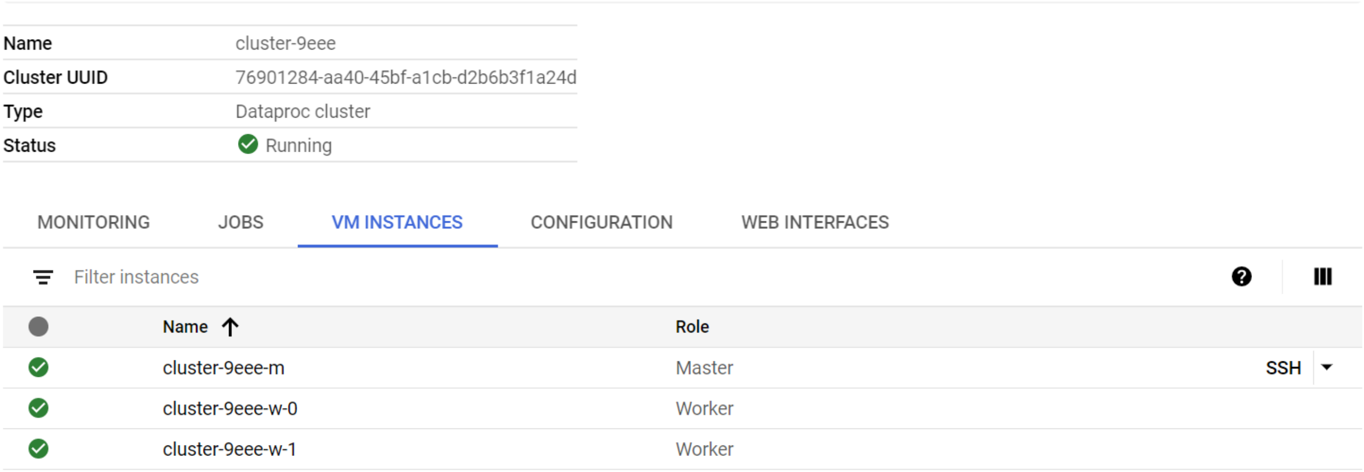
Cluster is created



□	Name	Region	Zone	Total worker nodes	Scheduled deletion	Cloud Storage staging bucket	Created	Status
□	✓ cluster-9eee	europe-west1	europe-west1-c	2	Off	dataproc-staging-europe-west1-895084807359-wjhqa4i1	18 Dec 2020, 15:33:18	Runnir



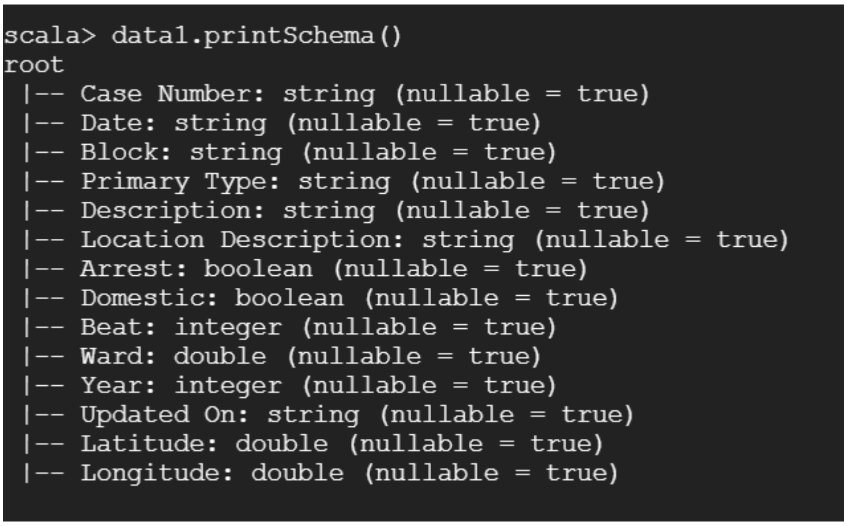
Name	cluster-9eee
Cluster UUID	76901284-aa40-45bf-a1cb-d2b6b3f1a24d
Type	Dataproc cluster
Status	✓ Running

●	Name	Role	
✓	cluster-9eee-m	Master	SSH
✓	cluster-9eee-w-0	Worker	
✓	cluster-9eee-w-1	Worker	

Data Loaded into Spark

val data1 = spark.read.options(Map("inferSchema"->"true","delimiter"->",","header"->"true")).csv("hdfs://cluster-9eee-m/user/soumi_mitra2/Crimes.csv")

data1.printSchema()

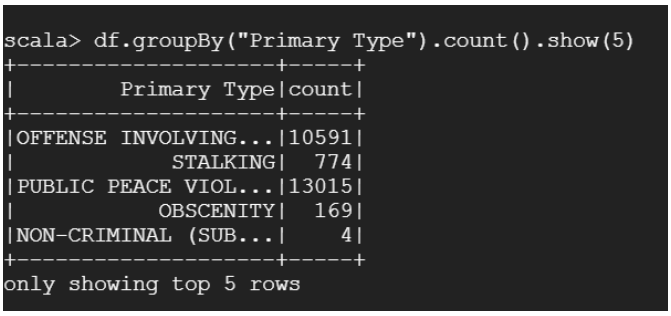


```
scala> data1.printSchema()
root
 |-- Case Number: string (nullable = true)
 |-- Date: string (nullable = true)
 |-- Block: string (nullable = true)
 |-- Primary Type: string (nullable = true)
 |-- Description: string (nullable = true)
 |-- Location Description: string (nullable = true)
 |-- Arrest: boolean (nullable = true)
 |-- Domestic: boolean (nullable = true)
 |-- Beat: integer (nullable = true)
 |-- Ward: double (nullable = true)
 |-- Year: integer (nullable = true)
 |-- Updated On: string (nullable = true)
 |-- Latitude: double (nullable = true)
 |-- Longitude: double (nullable = true)
```

val df= data1.toDF("Case Number","Date","Block","Primary Type","Description","Location Description","Arrest","Domestic","Beat","Ward","Year","Updated On","Latitude","Longitude")

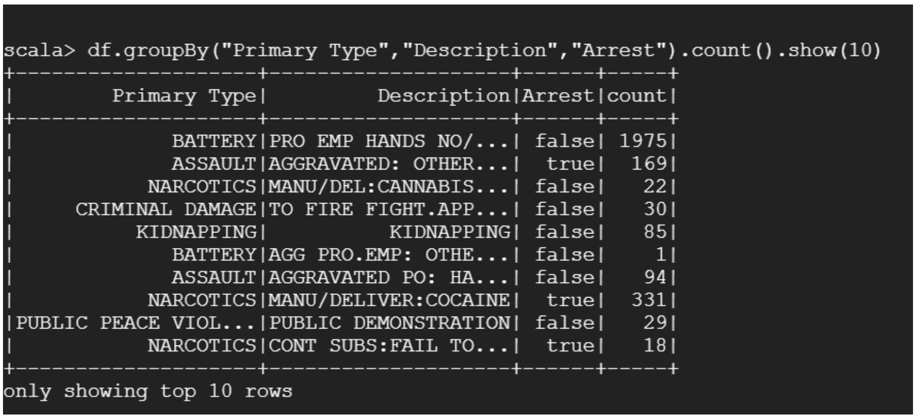
---- **Some statistics**

```
df.groupBy("Primary Type").count().show(5)
```



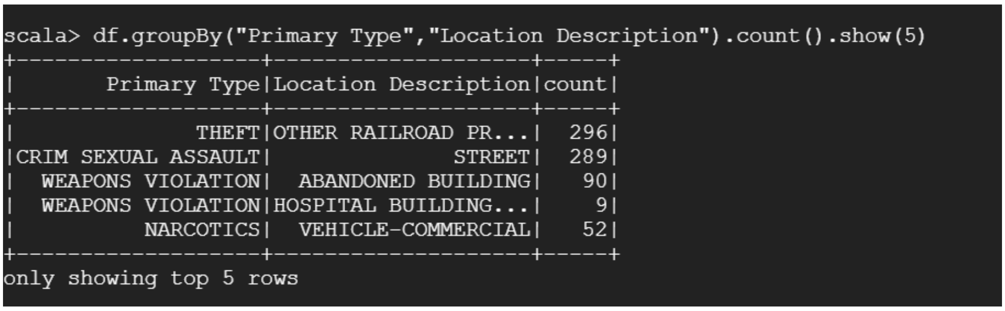
```
scala> df.groupBy("Primary Type").count().show(5)
+--------------------+-----+
|        Primary Type|count|
+--------------------+-----+
|OFFENSE INVOLVING...|10591|
|            STALKING|  774|
|PUBLIC PEACE VIOL...|13015|
|           OBSCENITY|  169|
|NON-CRIMINAL (SUB...|    4|
+--------------------+-----+
only showing top 5 rows
```

```
df.groupBy("Primary Type","Description","Arrest").count().show(10)
```



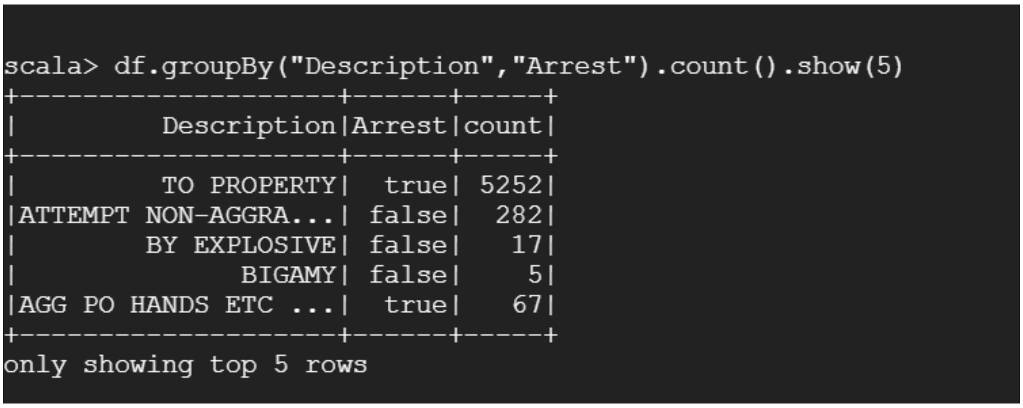
```
scala> df.groupBy("Primary Type","Description","Arrest").count().show(10)
+--------------------+--------------------+------+-----+
|        Primary Type|         Description|Arrest|count|
+--------------------+--------------------+------+-----+
|             BATTERY|PRO EMP HANDS NO/...| false| 1975|
|             ASSAULT|AGGRAVATED: OTHER...|  true|  169|
|           NARCOTICS|MANU/DEL:CANNABIS...| false|   22|
|     CRIMINAL DAMAGE|TO FIRE FIGHT.APP...| false|   30|
|          KIDNAPPING|          KIDNAPPING| false|   85|
|             BATTERY|AGG PRO.EMP: OTHE...| false|    1|
|             ASSAULT|AGGRAVATED PO: HA...| false|   94|
|           NARCOTICS|MANU/DELIVER:COCAINE|  true|  331|
|PUBLIC PEACE VIOL...|PUBLIC DEMONSTRATION| false|   29|
|           NARCOTICS|CONT SUBS:FAIL TO...|  true|   18|
+--------------------+--------------------+------+-----+
only showing top 10 rows
```

df.groupBy("Primary Type","Location Description").count().show(5)



```
scala> df.groupBy("Primary Type","Location Description").count().show(5)
+-------------------+--------------------+-----+
|       Primary Type|Location Description|count|
+-------------------+--------------------+-----+
|              THEFT|OTHER RAILROAD PR...|  296|
|CRIM SEXUAL ASSAULT|              STREET|  289|
|  WEAPONS VIOLATION|   ABANDONED BUILDING|   90|
|  WEAPONS VIOLATION|HOSPITAL BUILDING...|    9|
|          NARCOTICS|   VEHICLE-COMMERCIAL|   52|
+-------------------+--------------------+-----+
only showing top 5 rows
```

df.groupBy("Description","Arrest").count().show(5)



```
scala> df.groupBy("Description","Arrest").count().show(5)
+--------------------+------+-----+
|         Description|Arrest|count|
+--------------------+------+-----+
|         TO PROPERTY|  true| 5252|
|ATTEMPT NON-AGGRA...| false|  282|
|        BY EXPLOSIVE| false|   17|
|              BIGAMY| false|    5|
|AGG PO HANDS ETC ...|  true|   67|
+--------------------+------+-----+
only showing top 5 rows
```