Baltimore’daki 01/01/2011-06/18/2016 tarihleri arası suçların analizi

Databricks linki;

https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/5569809382116266/455069012075586/2421470568758723/latest.html

Ekran Görüntülerimiz ve yazdığımız kodların anlamları

URL aralıcığıyla internetten çekilen verileri String formatında tutuyoruz.

```
> def getUrlAsString(url: String): String = {
    val client = org.apache.http.impl.client.HttpClientBuilder.create().build()
    val request = new org.apache.http.client.methods.HttpGet(url)
    val response = client.execute(request)
    val handler = new org.apache.http.impl.client.BasicResponseHandler()
    handler.handleResponse(response).trim
  }
```

JSON verilerinin RDD aracılığıyla okunması işlemini yaptık.

```
> val namesRawData = sc.parallelize((0 until 1)).map {
      idx => getUrlAsString(s"https://data.baltimorecity.gov/api/views/wsfq-mvij/rows.json?accessType=DOWNLOAD")
  }
```

```
namesRawData: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[1071] at map at <console>:34
Command took 1.12 seconds -- by sammascolkesen@gmail.com at 26.12.2016 18:51:37 on baltimorecrime
```

SparkSQL'in DataFrame'ini kullanarak RDD'yi Spark DataFrame'ine çevirdik.

```
> val namesDataFrame = sqlContext.read.json(namesRawData)
```

```
▸ (1) Spark Jobs
namesDataFrame: org.apache.spark.sql.DataFrame = [data: array<array<string>>, meta: struct<view: struct<attribution: string,
re fields>>]
Command took 22.74 seconds -- by sammascolkesen@gmail.com at 26.12.2016 18:51:41 on baltimorecrime
```

SQL sorguları kullanabilmek için DataFrame'i tablo olarak kaydettik.

```
> namesDataFrame.createOrReplaceTempView("names_raw")
```

```
Command took 0.11 seconds -- by sammascolkesen@gmail.com at 26.12.2016 18:52:09 on baltimorecrime
```

Öncelikle DataFrame şemasını printSchema olarak kullanarak görüntüledik.

```
> namesDataFrame.printSchema
```

```
root
 |-- data: array (nullable = true)
 |    |-- element: array (containsNull = true)
 |    |    |-- element: string (containsNull = true)
 |-- meta: struct (nullable = true)
 |    |-- view: struct (nullable = true)
 |    |    |-- attribution: string (nullable = true)
 |    |    |-- attributionLink: string (nullable = true)
 |    |    |-- averageRating: long (nullable = true)
 |    |    |-- category: string (nullable = true)
 |    |    |-- columns: array (nullable = true)
 |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |-- cachedContents: struct (nullable = true)
 |    |    |    |    |    |-- average: string (nullable = true)
 |    |    |    |    |    |-- largest: string (nullable = true)
 |    |    |    |    |    |-- non_null: long (nullable = true)
 |    |    |    |    |    |-- null: long (nullable = true)
 |    |    |    |    |    |-- smallest: string (nullable = true)
 |    |    |    |    |    |-- sum: string (nullable = true)
 |    |    |    |    |    |-- top: array (nullable = true)
 |    |    |    |    |    |    |-- element: struct (containsNull = true)
```

Command took 0.13 seconds -- by sammascolkesen@gmail.com at 26.12.2016 18:52:14 on baltimorecrime

İlgili alanları data numaralarını 8'den itibaren başlatıp kısıtlamalarımızı buna göre yapacağız

```
> %sql
  select the_columns.fieldName, the_columns.position + 7 from names_raw lateral view explode(names_Raw.meta.view.columns) c as the_columns where
  the_columns.position > 0
```

▸ (2) Spark Jobs

fieldName	(the_columns.position AS `position` + CAST(7 AS BIGINT))
crimedate	8
crimetime	9
crimecode	10
location	11
description	12
inside_outside	13
weapon	14
post	15
district	16

Command took 17.41 seconds -- by sammascolkesen@gmail.com at 26.12.2016 18:52:20 on baltimorecrime

Sorguların kaç cpu'yu kullanarak çalışacağını belirttik.

```
> val partitionFactor = 4
```

```
partitionFactor: Int = 4
```

Command took 0.10 seconds -- by sammascolkesen@gmail.com at 26.12.2016 18:52:42 on baltimorecrime

Hangi verimizin hangi data sütununda bulunacağını belirttik.

```
> sql("""select the_data[8] as crimedate,the_data[9] as crimetime,the_data[10] as crimecode, the_data[14] as weapon,the_data[17] as neighborhood
  from names_raw lateral view explode(names_raw.data) d as the_data""")
  .repartition(sc.defaultParallelism * partitionFactor)
  .createOrReplaceTempView("baltimorecrime")
```

Command took 0.13 seconds -- by sammascolkesen@gmail.com at 26.12.2016 18:52:45 on baltimorecrime

Tablomuzun var olup olmadığını kontrol ediyoruz, eğer varsa hata çıkmaması için siliyoruz.

```
%sql
DROP TABLE IF EXISTS crimes_of_baltimore
```

▸ (1) Spark Jobs

OK

Command took 0.04 seconds -- by sammascolkesen@gmail.com at 26.12.2016 18:52:48 on baltimorecrime

Verilerimizle yeni tablomuzu oluşturuyoruz.

```
%sql CREATE TABLE crimes_of_baltimore AS
select * from baltimorecrime
```

▸ (2) Spark Jobs

OK

Command took 28.46 seconds -- by sammascolkesen@gmail.com at 26.12.2016 18:52:51 on baltimorecrime

Oluşturduğumuz tablomuzu önbelleğe kaydediyoruz.

```
%sql
CACHE TABLE crimes_of_baltimore
```

▸ (2) Spark Jobs

OK

Command took 1.81 seconds -- by sammascolkesen@gmail.com at 26.12.2016 18:53:33 on baltimorecrime

Verilerimizi gösteren SQL kodunuz yazarak görselleştiriyoruz.

```
%sql
select * from crimes_of_baltimore
```

▸ (1) Spark Jobs

crimedate	crimetime	crimecode	weapon	neighborhood
2016-12-17T00:00:00	13:45:00	3B	null	Oakenshawe
2016-12-17T00:00:00	02:35:00	3AF	FIREARM	Franklin Square
2016-12-16T00:00:00	1000	1K	KNIFE	Brooklyn
2016-12-16T00:00:00	19:43:00	4E	HANDS	Irvington
2016-12-15T00:00:00	05:30:00	4D	HANDS	Orchard Ridge
2016-12-15T00:00:00	14:05:00	6J	null	Riverside
2016-12-15T00:00:00	21:25:00	4C	OTHER	Kresson
2016-12-14T00:00:00	04:23:00	3CF	FIREARM	Downtown

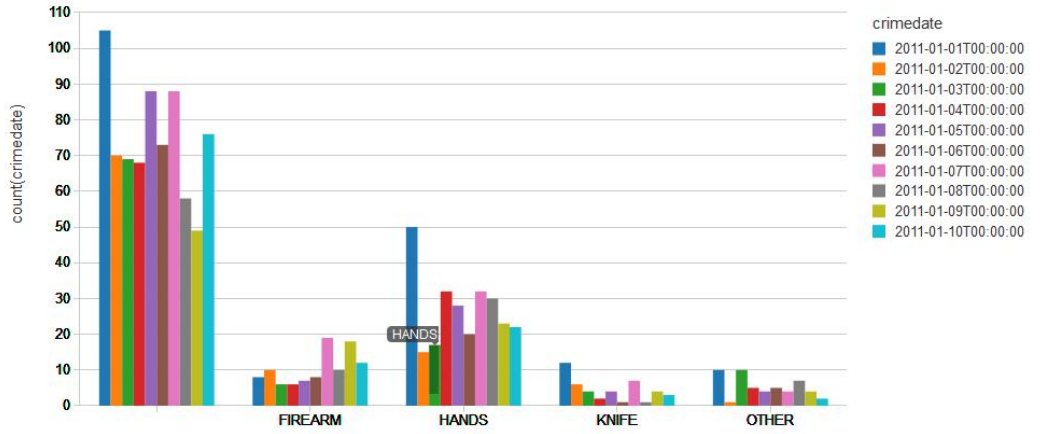
Showing the first 1000 rows.

Command took 0.14 seconds -- by sammascolkesen@gmail.com at 26.12.2016 18:53:37 on baltimorecrime

Send Feedback

Suç tipine göre o tarihte ne kadar suç işlendiği

```
%sql
select count(crimedate),crimedate,weapon
from crimes_of_baltimore
--where crimedate between '2013-11-02T00:00:00' and '2014-10-09T00:00:00'
group by crimedate,weapon
```



Only showing the first ten series.

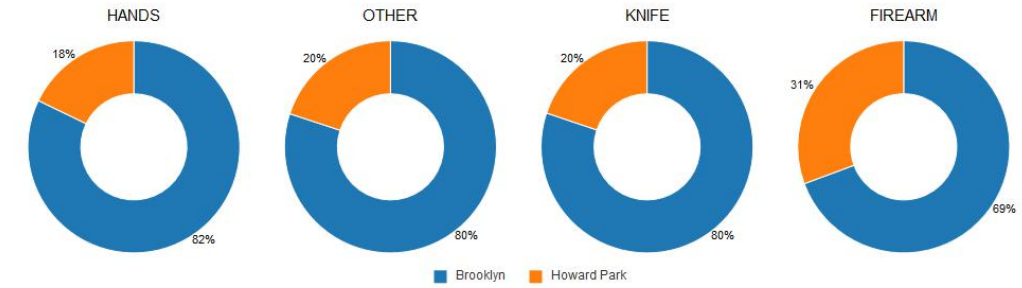
Plot Options...

Command took 2.15 seconds -- by sammascolkesen@gmail.com at 26.12.2016 19:09:07 on baltimorecrime

Brooklyn ve Howard Park taki işlenen suç tipi karşılaştırılması.

```
%sql
select count(*),neighborhood,weapon
from crimes_of_baltimore
where neighborhood='Brooklyn' or  neighborhood='Howard Park'
group by neighborhood,weapon
```

More than 4 pie charts. We will show only the first 4.



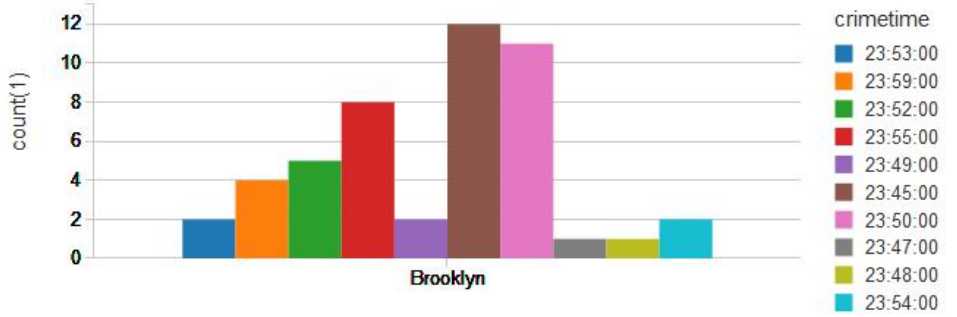
Plot Options...

Command took 0.37 seconds -- by sammascolkesen@gmail.com at 26.12.2016 19:14:55 on baltimorecrime

Brooklyn'de saat 23:45 ve 24:00 arası işlenen suç sayısı

```
%sql
select count(*),weapon,crimetime,neighborhood
from crimes_of_baltimore
where (crimetime between '23:45:00' and '24:00:00') and neighborhood='Brooklyn'
group by crimetime,weapon,neighborhood
--suç aleti belirtilmemiş olanlar belirlenememiş olanlar
```

▸ (2) Spark Jobs



Only showing the first ten series.

Plot Options...

Command took 0.35 seconds -- by sammascolkesen@gmail.com at 26.12.2016 19:32:11 on baltimorecrime

İstenen tarihteki suçların top 20 şehirlere oranı

```
%sql
select neighborhood
from crimes_of_baltimore
where crimedate='2011-01-02T00:00:00'
group by neighborhood  LIMIT 20
```

▸ (2) Spark Jobs