DATA 603

Assignment 3 : Using Apache Spark

Name : Akash Reddy Loka

Campus ID : LB04019

Installing pyspark library

In [1]:

```
import warnings
warnings.filterwarnings("ignore")
```

In [2]:

```
!pip install pyspark
```

```
Requirement already satisfied: pyspark in /opt/anaconda3/lib/python3.9/site-packages (3.3.2)
Requirement already satisfied: py4j==0.10.9.5 in /opt/anaconda3/lib/python3.9/site-packages (from pyspark) (0.10.9.5)
```

In [3]:

```
!pip install pyarrow
```

```
Requirement already satisfied: pyarrow in /opt/anaconda3/lib/python3.9/site-packages (11.0.0)
Requirement already satisfied: numpy>=1.16.6 in /opt/anaconda3/lib/python3.9/site-packages (from pyarrow) (1.21.5)
```

In [4]:

```
!pip install plotly
```

```
Requirement already satisfied: plotly in /opt/anaconda3/lib/python3.9/site-packages (5.6.0)
Requirement already satisfied: six in /opt/anaconda3/lib/python3.9/site-packages (from plotly) (1.16.0)
Requirement already satisfied: tenacity>=6.2.0 in /opt/anaconda3/lib/python3.9/site-packages (from plotly) (8.0.1)
```

1.Creating a Spark Session

In [5]:

```
from pyspark.sql import SparkSession
from pyspark.sql.types import StructType,StructField, StringType, IntegerType, DoubleType, BooleanType, DateType
spark = SparkSession.builder.appName("Chicago_crime_data").getOrCreate()
```

```
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/03/27 21:16:58 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

2.Defining the schema for loading the Chicago crime dataset

In [6]:

```
schema = StructType([StructField("ID",StringType(),True),
                     StructField("CaseNumber",StringType(),True),
                     StructField("Date",StringType(),True),
                     StructField("Block", StringType(), True),
                     StructField("IUCR", StringType(), True),
                     StructField("PrimaryType", StringType(), True),
                     StructField("Description", StringType(), True),
                     StructField("LocationDescription", StringType(), True),
                     StructField("Arrest", BooleanType(), True),
                     StructField("Domestic", BooleanType(), True),
                     StructField("Beat", StringType(), True),
                     StructField("District", StringType(), True),
                     StructField("Ward", StringType(), True),
                     StructField("CommunityArea", StringType(), True),
                     StructField("FBICode", StringType(), True ),
                     StructField("XCoordinate", DoubleType(), True),
                     StructField("YCoordinate", DoubleType(), True ),
                     StructField("Year", IntegerType(), True),
                     StructField("UpdatedOn", StringType(), True ),
                     StructField("Latitude", DoubleType(), True),
                     StructField("Longitude", DoubleType(), True),
                     StructField("Location", StringType(), True )
                     ])
```

3.Loading the Chicago crime data (you should get more than a million rows).

In [7]:
```
chicago_crime_data = spark.read.csv('Crimes_-_2001_to_Present.csv',
                          header = True,
                          schema = schema)
```

In [8]:
```
print('Chicago Crime dataset has',chicago_crime_data.count(),'rows')
```

```
[Stage 0:=====================================>                    (9 + 5) / 14]
Chicago Crime dataset has 7760248 rows
```

In [9]:
```
#printing first 5 rows
from IPython.core.display import HTML
display(HTML("<style>pre { white-space: pre !important; }</style>"))
chicago_crime_data.show(5,truncate = False)
```

```
23/03/27 21:17:10 WARN CSVHeaderChecker: CSV header does not conform to the schema.
 Header: ID, Case Number, Date, Block, IUCR, Primary Type, Description, Location Description, Arrest, Domestic, Beat, District, Ward, Community Area, FBI Code, X Coordinate, Y Coordina
 Schema: ID, CaseNumber, Date, Block, IUCR, PrimaryType, Description, LocationDescription, Arrest, Domestic, Beat, District, Ward, CommunityArea, FBICode, XCoordinate, YCoordinate, Yea
Expected: CaseNumber but found: Case Number
CSV file: file:///Users/akashreddyloka/Documents/UMBC/DS%20603/Assignments/Week%204/Akash_Pyspark_Assignment%20(1)/Crimes_-_2001_to_Present.csv
+--------+----------+----------------------+---------------------+----+-----------+-----------------------+-------------------+------+--------+----+--------+----+-------------+-------
|ID      |CaseNumber|Date                  |Block                |IUCR|PrimaryType|Description            |LocationDescription|Arrest|Domestic|Beat|District|Ward|CommunityArea|FBICode
+--------+----------+----------------------+---------------------+----+-----------+-----------------------+-------------------+------+--------+----+--------+----+-------------+-------
|10224738|HY411648  |09/05/2015 01:30:00 PM|043XX S WOOD ST      |0486|BATTERY    |DOMESTIC BATTERY SIMPLE|RESIDENCE          |false |true    |0924|009     |12  |61           |08B
|10224739|HY411615  |09/04/2015 11:30:00 AM|008XX N CENTRAL AVE  |0870|THEFT      |POCKET-PICKING         |CTA BUS            |false |false   |1511|015     |29  |25           |06
|11646166|JC213529  |09/01/2018 12:01:00 AM|082XX S INGLESIDE AVE|0810|THEFT      |OVER $500              |RESIDENCE          |false |true    |0631|006     |8   |44           |06
|10224740|HY411595  |09/05/2015 12:45:00 PM|035XX W BARRY AVE    |2023|NARCOTICS  |POSS: HEROIN(BRN/TAN)  |SIDEWALK           |true  |false   |1412|014     |35  |21           |18
|10224741|HY411610  |09/05/2015 01:00:00 PM|0000X N LARAMIE AVE  |0560|ASSAULT    |SIMPLE                 |APARTMENT          |false |true    |1522|015     |28  |25           |08A
+--------+----------+----------------------+---------------------+----+-----------+-----------------------+-------------------+------+--------+----+--------+----+-------------+-------
only showing top 5 rows
```

4.Clean the data:

4.a Remove all null values.

In [10]:
```
chicago_crime_data_null_removed = chicago_crime_data.dropna()
```

In [11]:
```
print('Chicago Crime dataset has',chicago_crime_data_null_removed.count(),\
      'rows after removing records which has null values')
```

```
23/03/27 21:17:11 WARN CSVHeaderChecker: CSV header does not conform to the schema.
 Header: ID, Case Number, Date, Block, IUCR, Primary Type, Description, Location Description, Arrest, Domestic, Beat, District, Ward, Community Area, FBI Code, X Coordinate, Y Coordina
 Schema: ID, CaseNumber, Date, Block, IUCR, PrimaryType, Description, LocationDescription, Arrest, Domestic, Beat, District, Ward, CommunityArea, FBICode, XCoordinate, YCoordinate, Yea
Expected: CaseNumber but found: Case Number
CSV file: file:///Users/akashreddyloka/Documents/UMBC/DS%20603/Assignments/Week%204/Akash_Pyspark_Assignment%20(1)/Crimes_-_2001_to_Present.csv
[Stage 4:================================================>        (12 + 2) / 14]
Chicago Crime dataset has 7061244 rows after removing records which has null values
```

4.b Change 'Date' column data type

In [12]:

```
chicago_crime_data_null_removed.printSchema()
```

```
root
 |-- ID: string (nullable = true)
 |-- CaseNumber: string (nullable = true)
 |-- Date: string (nullable = true)
 |-- Block: string (nullable = true)
 |-- IUCR: string (nullable = true)
 |-- PrimaryType: string (nullable = true)
 |-- Description: string (nullable = true)
 |-- LocationDescription: string (nullable = true)
 |-- Arrest: boolean (nullable = true)
 |-- Domestic: boolean (nullable = true)
 |-- Beat: string (nullable = true)
 |-- District: string (nullable = true)
 |-- Ward: string (nullable = true)
 |-- CommunityArea: string (nullable = true)
 |-- FBICode: string (nullable = true)
 |-- XCoordinate: double (nullable = true)
 |-- YCoordinate: double (nullable = true)
 |-- Year: integer (nullable = true)
 |-- UpdatedOn: string (nullable = true)
 |-- Latitude: double (nullable = true)
 |-- Longitude: double (nullable = true)
 |-- Location: string (nullable = true)
```

It can be observed that the Date column is of string datatype.

In [13]:

```
#converting the Date column datatype from String to timestamp in 24 Hr format
from pyspark.sql.functions import to_timestamp
chicago_crime_datetype_modified = chicago_crime_data_null_removed.withColumn("Date",\
                                  to_timestamp('Date',"MM/dd/yyyy hh:mm:ss a"))
```

In [14]:

```
chicago_crime_datetype_modified.printSchema()
```

```
root
 |-- ID: string (nullable = true)
 |-- CaseNumber: string (nullable = true)
 |-- Date: timestamp (nullable = true)
 |-- Block: string (nullable = true)
 |-- IUCR: string (nullable = true)
 |-- PrimaryType: string (nullable = true)
 |-- Description: string (nullable = true)
 |-- LocationDescription: string (nullable = true)
 |-- Arrest: boolean (nullable = true)
 |-- Domestic: boolean (nullable = true)
 |-- Beat: string (nullable = true)
 |-- District: string (nullable = true)
 |-- Ward: string (nullable = true)
 |-- CommunityArea: string (nullable = true)
 |-- FBICode: string (nullable = true)
 |-- XCoordinate: double (nullable = true)
 |-- YCoordinate: double (nullable = true)
 |-- Year: integer (nullable = true)
 |-- UpdatedOn: string (nullable = true)
 |-- Latitude: double (nullable = true)
 |-- Longitude: double (nullable = true)
 |-- Location: string (nullable = true)
```

It can be observed that the Date column datatype has been changed to timestamp.

In [15]:

```python
#printing first 5 rows to check for the modified datatype in Date column
from IPython.core.display import HTML
display(HTML("<style>pre { white-space: pre !important; }</style>"))
chicago_crime_datetype_modified.show(5,truncate=False)
```

```
23/03/27 21:17:20 WARN CSVHeaderChecker: CSV header does not conform to the schema.
 Header: ID, Case Number, Date, Block, IUCR, Primary Type, Description, Location Description, Arrest, Domestic, Beat, District, Ward, Community Area, FBI Code, X Coordinate, Y Coordina
 Schema: ID, CaseNumber, Date, Block, IUCR, PrimaryType, Description, LocationDescription, Arrest, Domestic, Beat, District, Ward, CommunityArea, FBICode, XCoordinate, YCoordinate, Yea
Expected: CaseNumber but found: Case Number
CSV file: file:///Users/akashreddyloka/Documents/UMBC/DS%20603/Assignments/Week%204/Akash_Pyspark_Assignment%20(1)/Crimes_-_2001_to_Present.csv
+--------+----------+-------------------+-------------------+----+-----------+-----------------------+-------------------+------+--------+----+--------+----+-------------+-------+-----
|ID      |CaseNumber|Date               |Block              |IUCR|PrimaryType|Description            |LocationDescription|Arrest|Domestic|Beat|District|Ward|CommunityArea|FBICode|XCoo
+--------+----------+-------------------+-------------------+----+-----------+-----------------------+-------------------+------+--------+----+--------+----+-------------+-------+-----
|10224738|HY411648  |2015-09-05 13:30:00|043XX S WOOD ST    |0486|BATTERY    |DOMESTIC BATTERY SIMPLE|RESIDENCE          |false |true    |0924|009     |12  |61           |08B    |1165
|10224739|HY411615  |2015-09-04 11:30:00|008XX N CENTRAL AVE|0870|THEFT      |POCKET-PICKING         |CTA BUS            |false |false   |1511|015     |29  |25           |06     |1138
|10224740|HY411595  |2015-09-05 12:45:00|035XX W BARRY AVE  |2023|NARCOTICS  |POSS: HEROIN(BRN/TAN)  |SIDEWALK           |true  |false   |1412|014     |35  |21           |18     |1152
|10224741|HY411610  |2015-09-05 13:00:00|0000X N LARAMIE AVE|0560|ASSAULT    |SIMPLE                 |APARTMENT          |false |true    |1522|015     |28  |25           |08A    |1141
|10224742|HY411435  |2015-09-05 10:55:00|082XX S LOOMIS BLVD|0610|BURGLARY   |FORCIBLE ENTRY         |RESIDENCE          |false |false   |0614|006     |21  |71           |05     |1168
+--------+----------+-------------------+-------------------+----+-----------+-----------------------+-------------------+------+--------+----+--------+----+-------------+-------+-----
only showing top 5 rows
```

5. Filter the data for last ten years.

In [16]:

```python
chicago_crime_datetype_modified.createOrReplaceTempView("temp_view")
chicago_crime_10_years = spark.sql('select * from temp_view where Year >=2014')
```

6. Remove all the records with the following crime types: 'NON-CRIMINAL (SUBJECT SPECIFIED)' 'OTHER OFFENSE' 'STALKING' 'NON - CRIMINAL' 'ARSON'

In [17]:

```python
chicago_crime_10_years.createOrReplaceTempView("temp_view_2")
chicago_crime_crimetypes_removed = spark.sql("select * from temp_view_2 where PrimaryType not in " +
         "('NON-CRIMINAL (SUBJECT SPECIFIED)','OTHER OFFENSE', 'STALKING','NON - CRIMINAL','ARSON') ")
```

7. Merge the similar crime types.

For example, change 'Primary Type' of cases that have 'Primary Type' as ‘SEX OFFENSE’ or ‘PROSTITUTION’ such that they should have the same 'Primary Type'.

In [18]:

```
from pyspark.sql.functions import when
crime_data_crimetype_merged = chicago_crime_crimetypes_removed.withColumn("PrimaryType", \
                       when(chicago_crime_crimetypes_removed.PrimaryType == "PROSTITUTION","SEX OFFENSE").\
                       otherwise(chicago_crime_crimetypes_removed.PrimaryType))
```

8. Analyze the data and present results:

8.a Show year-wise trend of the crime for last ten years.

In [19]:

```
crime_data_crimetype_merged.createOrReplaceTempView("temp_view_3")
crime_trend_yearwise = spark.sql("select year,count(*) as crime_counts " +
                                 "from temp_view_3 group by year order by year asc")
```

In [20]:

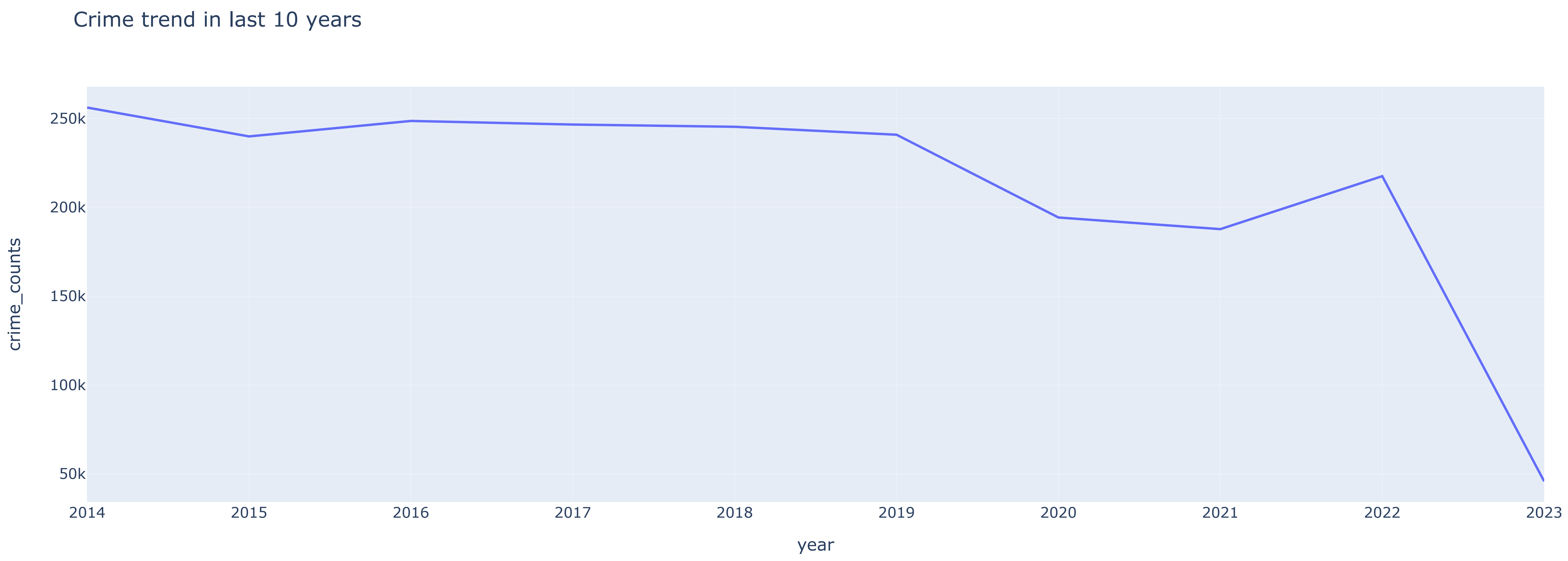
```
import os
os.environ["PYARROW_IGNORE_TIMEZONE"] = "1"
```

In [21]:

```
crime_trend_yearwise.pandas_api().plot.line(x='year',y='crime_counts',title='Crime trend in last 10 years')
```

```
23/03/27 21:17:22 WARN CSVHeaderChecker: CSV header does not conform to the schema.
 Header: ID, Case Number, Date, Block, IUCR, Primary Type, Description, Location Description, Arrest, Domestic, Beat, District, Ward, Community Area, FBI Code, X Coordinate, Y Coordina
 Schema: ID, CaseNumber, Date, Block, IUCR, PrimaryType, Description, LocationDescription, Arrest, Domestic, Beat, District, Ward, CommunityArea, FBICode, XCoordinate, YCoordinate, Yea
Expected: CaseNumber but found: Case Number
CSV file: file:///Users/akashreddyloka/Documents/UMBC/DS%20603/Assignments/Week%204/Akash_Pyspark_Assignment%20(1)/Crimes_-_2001_to_Present.csv
```

```
23/03/27 21:17:30 WARN CSVHeaderChecker: CSV header does not conform to the schema.
 Header: ID, Case Number, Date, Block, IUCR, Primary Type, Description, Location Description, Arrest, Domestic, Beat, District, Ward, Community Area, FBI Code, X Coordinate, Y Coordina
 Schema: ID, CaseNumber, Date, Block, IUCR, PrimaryType, Description, LocationDescription, Arrest, Domestic, Beat, District, Ward, CommunityArea, FBICode, XCoordinate, YCoordinate, Yea
Expected: CaseNumber but found: Case Number
CSV file: file:///Users/akashreddyloka/Documents/UMBC/DS%20603/Assignments/Week%204/Akash_Pyspark_Assignment%20(1)/Crimes_-_2001_to_Present.csv
```


Crime trend in last 10 years

Observations:

It can be observed from the above line chart that

1.Year 2014 has highest number of crimes.

2.Year 2023 has the lowest number of crimes.

3.Crime Rate has been reduced drastically from the year 2022 to 2023.

8.b Find out at which hour of the day crime is highest.

```
In [22]: spark.sql("select Hour(Date) as Hour, Count(*) as Crime_counts from temp_view_3 " +
                   "group by Hour order by Crime_counts desc").show()
```

```
23/03/27 21:17:39 WARN CSVHeaderChecker: CSV header does not conform to the schema.
 Header: ID, Case Number, Date, Block, IUCR, Primary Type, Description, Location Description, Arrest, Domestic, Beat, District, Ward, Community Area, FBI Code, X Coordinate, Y Coordina
 Schema: ID, CaseNumber, Date, Block, IUCR, PrimaryType, Description, LocationDescription, Arrest, Domestic, Beat, District, Ward, CommunityArea, FBICode, XCoordinate, YCoordinate, Yea
Expected: CaseNumber but found: Case Number
CSV file: file:///Users/akashreddyloka/Documents/UMBC/DS%20603/Assignments/Week%204/Akash_Pyspark_Assignment%20(1)/Crimes_-_2001_to_Present.csv
[Stage 22:=====================================================>   (13 + 1) / 14]
```

```
+----+------------+
|Hour|Crime_counts|
+----+------------+
|  12|      123290|
|  18|      119921|
|  19|      118948|
|  15|      115906|
|  17|      115116|
|   0|      114445|
|  20|      113716|
|  16|      112852|
|  21|      106141|
|  14|      106109|
|  22|      104483|
|  13|      100966|
|  11|       95240|
|   9|       92097|
|  10|       91898|
|  23|       88292|
|   8|       71083|
|   1|       66656|
|   2|       57979|
|   7|       50595|
+----+------------+
only showing top 20 rows
```

It is found that, at 12th hour the crime rate is highest.

8.c Find top ten crimes and present them as a bar chart.

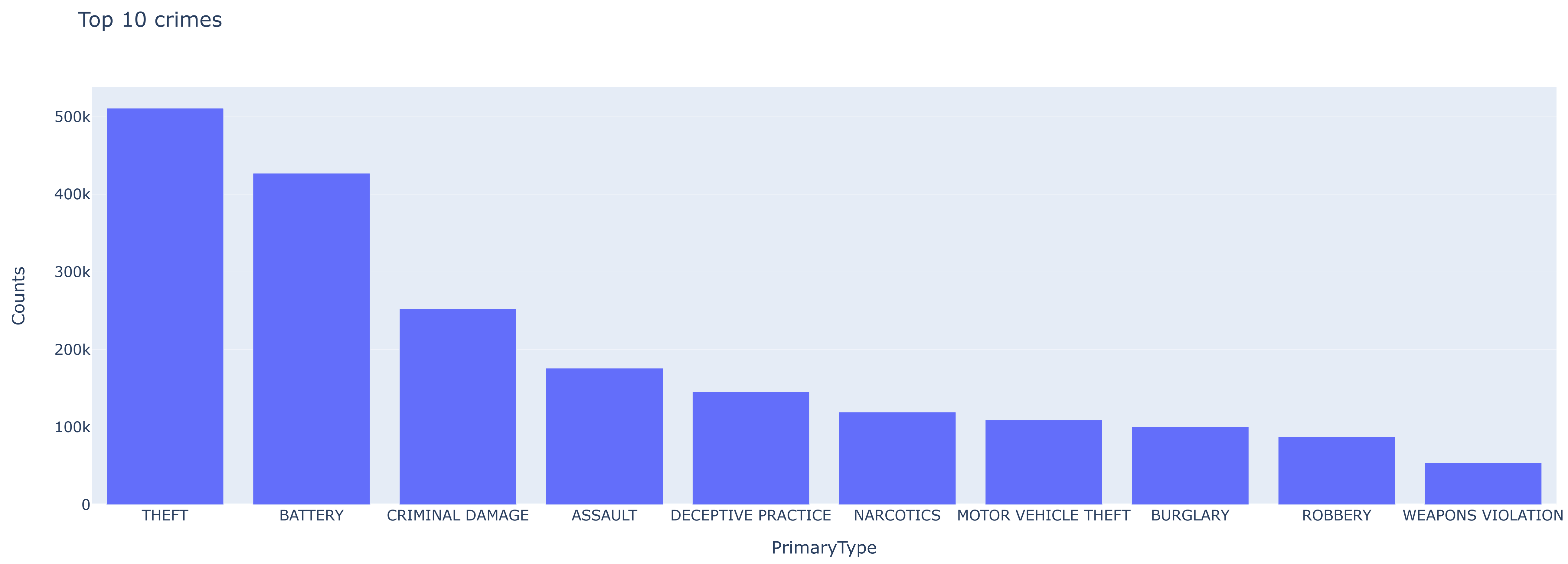
In [23]:

```
df_top10_crimes = spark.sql("select PrimaryType,Count(*) as Counts "+
                            "from temp_view_3 group by PrimaryType order by Counts desc limit 10")
```

In [24]:

```
df_top10_crimes.pandas_api().plot(kind='bar',x='PrimaryType',y='Counts',title='Top 10 crimes')
```

```
[Stage 25:>                                                         (0 + 8) / 14]
23/03/27 21:17:50 WARN CSVHeaderChecker: CSV header does not conform to the schema.
 Header: ID, Case Number, Date, Block, IUCR, Primary Type, Description, Location Description, Arrest, Domestic, Beat, District, Ward, Community Area, FBI Code, X Coordinate, Y Coordina
 Schema: ID, CaseNumber, Date, Block, IUCR, PrimaryType, Description, LocationDescription, Arrest, Domestic, Beat, District, Ward, CommunityArea, FBICode, XCoordinate, YCoordinate, Yea
Expected: CaseNumber but found: Case Number
CSV file: file:///Users/akashreddyloka/Documents/UMBC/DS%20603/Assignments/Week%204/Akash_Pyspark_Assignment%20(1)/Crimes_-_2001_to_Present.csv
```



Observations:

It can be observed from the above Bar chart that

1.Theft crimes are highest in number.

2.Weapons violation crimes are lowest in number.