**Pyspark Assignment1**

**Task 1:**

**Crime Data Analysis**

**Dataset: crime\_incidents\_2013.csv**

[mavricbdhoct07@ip-10-1-1-204 ~]$ hdfs dfs -put crime\_incidents\_2013.csv

[mavricbdhoct07@ip-10-1-1-204 ~]$ hdfs dfs -ls

A screenshot of a computer

Description automatically generated

crimeDF=spark.read.option("header",True).option("inferschema",True).csv("crime\_incidents\_2013.csv")

crimeDF.createOrReplaceTempView("crime")

crimeDF.printSchema()

A screenshot of a computer

Description automatically generated

crimeDF.show()

A screenshot of a computer

Description automatically generated

**7. How many offenses were there in 2013 ?**

spark.sql("SELECT count(distinct OFFENSE) from crime").show()

A screenshot of a computer

Description automatically generated

**8. How many crimes were committed in each offense?**

spark.sql("SELECT OFFENSE ,count(CCN) from crime group by OFFENSE").show()

A screenshot of a computer

Description automatically generated

**9. How many different methods were used (And their count) in offense**

**“Homicide”**

spark.sql("SELECT METHOD ,count(METHOD) from crime where OFFENSE = 'HOMICIDE' ").show()

A screenshot of a computer

Description automatically generated

**10. Which shift had the maximum crimes?**

spark.sql("SELECT SHIFT ,max(CCN) from crime group by SHIFT order by 2 desc").show(1)

A screenshot of a computer

Description automatically generated

spark.sql("SELECT SHIFT ,max(CCN) from crime group by SHIFT").show()

A screenshot of a computer

Description automatically generated

spark.sql("SELECT SHIFT ,max(CCN) from crime group by SHIFT order by 2 desc").show()

A screenshot of a computer

Description automatically generated