INCEPTEZ TECHNOLOGIES Spark SQL WORKOUTS

Prerequisite:
Copy the sfpd.csv into /home/hduser path.

Use DataFrames to load data into Spark using CSV data

```
from pyspark.sql import SQLContext, Row
import pyspark.sql.functions as func
sqlContext = SQLContext(sc)

#Create input RDD
sfpdRDD = sc.textFile("file:///home/hduser/sfpd.csv").map(lambda inc: inc.split(","))

# Infer the schema, and register the DataFrame as a table.
sfpdSchema=sfpdRDD.map(lambda inc: Row(incidentnum=inc[0],category=inc[1],
description=inc[2],dayofweek=inc[3],date=inc[4],time=inc[5],pddistrict=inc[6],resolution=inc[7],address
=inc[8],X=float(inc[9]),Y=float(inc[10]), pdid=inc[11]))

sfpdDF=sqlContext.createDataFrame(sfpdSchema)

sfpdDF.registerTempTable("sfpd")

#1. Top 5 Districts
incByDist = sfpdDF.groupBy("pddistrict").count().sort(func.desc("count"))
incByDist.show(5)

topByDistSQL = sqlContext.sql("SELECT pddistrict, count(incidentnum) AS inccount FROM sfpd GROUP
BY pddistrict ORDER BY inccount DESC LIMIT 5")
topByDistSQL.show()

#2. What are the top ten resolutions?
top10Res = sfpdDF.groupBy("resolution").count().sort(func.desc("count"))
top10Res.show(10)
top10ResSQL = sqlContext.sql("SELECT resolution, count(incidentnum) AS inccount FROM sfpd GROUP
BY resolution ORDER BY inccount DESC LIMIT 10")
top10ResSQL.show()

#3. Top 3 categories
```

```
top3Cat = sfpdDF.groupBy("category").count().sort(func.desc("count"))
top3Cat.show(3)
top3CatSQL=sqlContext.sql("SELECT category, count(incidentnum) AS inccount FROM sfpd GROUP BY category ORDER BY inccount DESC LIMIT 3")
top3CatSQL.show()
```

User Defined Functions

```
#UDF with SQL
#You can use registerFunction to register a Lambda function.

# Register the function as a udf
sqlContext.registerFunction("getyear",lambda x:x[-2:])

# Count inc by year
incyearSQL=sqlContext.sql("SELECT getyear(date), count(incidentnum) AS countbyyear FROM sfpd GROUP BY getyear(date) ORDER BY countbyyear DESC")
incyearSQL.show()

#Category, resolution and address of reported incidents in 2014
inc2014 = sqlContext.sql("SELECT category,address,resolution, date FROM sfpd WHERE getyear(date)='14'")
inc2014.show()
# Can also use collect()

#Vandalism only in 2014 with address, resolution and category
van2015 = sqlContext.sql("SELECT category,address,resolution, date FROM sfpd WHERE getyear(date)='15' AND category='VANDALISM'")
van2015.show()
van2015.count()
```