EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

TECNOLÓGICO NACIONAL DE MÉXICO

INSTITUTO TECNOLÓGICO DE TIJUANA
SUBDIRECCIÓN ACADÉMICA
Departamento de Sistemas y Computación
EXAMEN

Carrera: Ingeniería En Sistemas Computacionales/ Tecnologías de la información/ Informática		Período: Agosto-Diciembre 2021
Materia: Datos Masivos	Grupo: BDD-1704	Salón:
Unidad (es) a evaluar: Unidad 1	Tipo de examen:Practico	Fecha:
Catedrático: Jose Christian Romero Hernandez	Firma del maestro:	Calificación:

Alumno: César Alejandro Velázquez Farrera _____ No Control: 17212937_____

Instrucciones

Responder las siguientes preguntas con Spark DataFrames y Scala utilizando el “CSV” Netflix_2011_2016.csv que se encuentra en la carpeta de spark.dataframes

1. Comienza una simple sesión en Spark

```
import org.apache.spark.sql.SparkSession
val session = SparkSession.builder().getOrCreate
```

2. Cargue el archivo Netflix Stock CSV, haga que spark infiera los tipos de datos

```
val df_netflix = session.read.option("header",
"true").option("inferSchema", true).csv("Netflix_2011_2016.csv")
```

3. ¿Cuáles son los nombres de las columnas?

```
df_netflix.columns
```

4. ¿Cómo es el esquema?

```
df_netflix.printSchema()
```

5. Imprime las primeras 5 columnas

```
df_netflix.head(5)
```

6. Usa describe() para aprender sobre el DataFrame

```
df_netlfix.describe().show
```

7. Crea un nuevo data frame con una columna nueva llamada “HV Ratio” que es la relación que existe entre el precio de la columna “High” frente a la columna “Volumen” de acciones negociadas por un día. *Hint es una operación.*

```
val df_netflix2 = df_netflix.withColumn("HV Ratio",
df_netflix("High")/df_netflix("Volume"))
```

8. ¿Qué día tuvo el pico más alto en la columna “Open”?

```
df_netflix.select(mean("Open")).show()
```

9. ¿Cuál es el significado de la columna Cerrar “Close” en el contexto de información financiera?

```
df_netflix.select(mean("Close")).show()
```

10. ¿Cuál es el máximo y mínimo de la columna “Volumen”?

```
df_netflix.select(max("Volume")).show()
df_netflix.select(min("Volume")).show()
```

11. Con sintaxis Scala/Spark $ conteste lo siguiente:
    a. ¿Cuántos días fue la columna “Close” inferior a $600?

```
val Day = df_netflix.where($"Close" < 600).count()
```

    b. ¿Qué porcentaje del tiempo fue la columna “High” mayor que $500?

```
val Day = df_netflix.where($"High" > 500).count().toFloat
```

    c. ¿Cuál es la correlación de Pearson entre la columna “High” y la columna “Volumen”?

```
df_netflix.select(corr("High", "Volume")).show()
```

    d. ¿Cuál es el máximo de la columna “High” por año?

```
df_netflix.groupBy(year($"Date")).max("High").show()
```

    e. ¿Cuál es el promedio de la columna “Close” para cada mes del calendario?

```
val df_netflix3 = df_netflix.groupBy(year($"Date"),
month($"Date")).mean("Close"). toDF("Year","Month","Mean")
df_netflix3.orderBy($"Year",$"Month").show()
```

Conclusion

En este examen pudimos observar lo eficiente que es el programa Apache Spark para la consulta y análisis, tratándose de un CSV como lo son los datos del archivo “Netflix_2011_2016.csv” como lo son la duración, volumen y la fecha de los datos que se encuentran dentro de este mismo.