Tecnológico Nacional de México
INSTITUTO TECNOLÓGICO CAMPUS TIJUANA
ING. EN SISTEMAS COMPUTACIONALES

Subdirección Académica
Departamento de Sistemas y Computación
BDD-1704 TI9A - 6:00pm-7:00pm

ASIGNATURA:
Datos Masivos

SEMESTRE:
Septiembre- Enero 2020

Examen:
Práctica Evaluatoria

MAESTRO:
JOSE CHRISTIAN ROMERO HERNANDEZ

Equipo:
Marco Antonio Rodriguez Medrano
Aide Ceballos Bobadilla

24/10/2020

Introducción:

buenos días, tardes o noches querido profesor, se nos pidió a mi, el alumno: Marco Antonio Rodriguez Medrano y a mi compañera de equipo la alumna: Aide Ceballos Bobadilla, hacer una práctica evaluatoria la cual consiste en aplicar nuestros conocimientos sobre DataFrames vistos en clase, pero por la actual pandemia que nos tiene aislados.

Tuvimos que trabajar remotamente con la ayuda del software TeamViewer y la plataforma meet , los cuales nos ayudaron a primero conectarnos vía remota y trabajar los dos en una misma computadora y el segundo a tener un diálogo (pair coding), el cual consiste en que un miembro del equipo tomará el rol de programador y el segundo tomará el rol de asesor del programador y después de un tiempo se cambiarán los roles los miembros del equipo.

DESARROLLO:

//antes de siquiera trabajar con dataframes debemos de forma obligatoria importar la sesion apache spark

```
//1)importamos la sesion apache spark
import org.apache.spark.sql.SparkSession
```

//después de cargar la sesion creamos la variable spark y le cargaremos a la misma la sesión

```
//2) creamos la variable spark
val spark = SparkSession.builder().getOrCreate()
```

//aqui creamos la variable df y le asignamos la variable spark previamente cargada y por ultimo cargamos el archivo csv

```
/*creamos la variable df y le asignamos la variable spark y cargamos el archivo
Netflix_2011_2016.csv*/
val df = spark.read.option("header",
"true").option("inferSchema","true")csv("C:/Users/aide0/OneDrive/Escritorio/Practica
Evaluatoria/Practica-Evaluatoria/Netflix_2011_2016.csv")
```

//a continuación creamos los comandos utilizando Dataframes

//muestra los nombres de las columnas de nuestro archivo .csv

```
//3)Nombres de las columnas
df.columns
```

//En este punto se muestra el esquema que correspondiente a el archivo .csv

```
//4)Esquema
```

```
df.printSchema ()
```

// Con el comando que se muestra a continuación se despliegan los primeros 5 registros del archivo .csv

```
//5)Primeras 5 columnas
df.take(5)
```

//con este comando se describe el maximo,minimo,media, desviación estándar y cuenta los valores de las columnas del archivo csv

```
// 6)Use describe () to learn about the DataFrame.
df.describe().show
```

// Se creó un dataframe con una columna llamada “HV Ratio” la cual es la relación entre el las columnas “High” y “Volume”

```
/*7)Crea un nuevo dataframe con una columna nueva llamada “HV Ratio” que es la relación
entre el precio de la columna “High” frente a la columna “Volume” de
acciones negociadas por un día. (Hint: Es una operación*/

val df2 = df.withColumn("HV_Ratio", df("High")/df("Volume"))
 df2.show
```

// Con los siguientes comandos se muestra los días por Mes y Semana de los días más altos de la columna “Close” los cuales fueron: Mes Dia 13 y Semana Dia 2

```
 //8)¿que dia tuvo el pico mas alto en la columna"close"?
 df.groupBy(dayofmonth(df("date")).alias("Day")).max("High").sort(asc("Day")).show()
 df.groupBy(dayofweek(df("date")).alias("Day")).max("High").sort(asc("Day")).show()
```

//**aqui como menciona la pregunta describimos con nuestras propias palabras lo que para nosotros significa la columna close .**

```
 /*9)Escribe con tus propias palabras en un comentario de tu código. ¿Cuál es el significado
de la columna Cerrar “Close”?*/

//(Marco)close en la bolsa de valores de netflix del 2011 al 2016, es el valor total de las
acciones con las que cerró ese día.
//(Aide) La columna close se refiere a la comparación de cuál fue el valor con el que
cerraros las acciones de Netflix.
```

//**aquí únicamente mostramos el valor minimo y maximo que contiene la columna volume**

```
//10)¿Cuál es el máximo y mínimo de la columna “Volume”?
df.select(max("Volume"),min("Volume")).show()
```

//**aquí se nos indica que debemos utilizar la sintaxis de Scala para resolver los siguientes problemas y ademas crear otro dataframe**

```
/*11)Con Sintaxis Scala/Spark $ conteste los siguiente:
◦ Hint: Basicamente muy parecido a la session de dates, tendran que crear otro
dataframe para contestar algunos de los incisos.*/
```

//Se creó una nuevo dataframe con el nombre “ejer_11” para contestar los incisos siguientes

```
val df3 = df.withColumn("Resultados ejer_11", df("High")/df("Volume")/df("Close"))
 df3.show
```

// Para saber cuántos fueron los días inferiores a 600 se utilizó el siguiente comando, el cual nos dio como resultado 1218

```
 //11a)a. ¿Cuántos días fue la columna “Close” inferior a $ 600?
val  preciomenor = df3.filter($"Close" < 600).count
```

//aquí indicamos el porcentaje de la columna high en donde fue mayor a 500

```
//11b. ¿Qué porcentaje del tiempo fue la columna “High” mayor que $ 500?
val tiempo = df3.filter($"High" > 500).count()
val tiempo1 = tiempo * .100
```

//Aqui se expresa la relacion Person entre las columnas “High” y “Volumen”

```
//11c. ¿Cuál es la correlación de Pearson entre columna “High” y la columna “Volumen”?
df3.select(corr("High", "Volume").alias("correlacion")).show()
```

//aqui utilizando groupBy indicamos el valor maximo por añi de la columna high

```
//11d. ¿Cuál es el máximo de la columna “High” por año?
df3.groupBy(year(df("Date")).alias("Year")).max("High").sort(asc("Year")).show()
```

//aquí indicamos el promedio de la columna close en cada mes del año

```
//11e. ¿Cuál es el promedio de columna “Close” para cada mes del calendario?
df3.groupBy(month(df("Date")).alias("Month")).avg("Close").sort(asc("Month")).show()
```

conclusión:

(Marco Antonio).- Los DataFrames en spark como lo explicó el profesor son colecciones distribuidas de datos, organizadas en filas y columnas, cada columna de un DataFrame tiene un nombre y un tipo asociado y permiten el procesamiento de grandes cantidades de datos o como es conocido Big Data.

Esto es muy bueno cuando manejamos una cantidad de datos considerables, ya que nos ayuda a optimizar los recursos de la computadora ya que si no utilizamos DataFrames para limpiar los datos, las consultas, búsqueda, y entre otras operaciones tomarían mucho tiempo y espacio de memoria.

(Aide Ceballos).- A lo largo de la clase aprendimos que es un DataFrame y los BigData, cuales son sus usos e importancia, con la Práctica Evaluatoria que acabamos de desarrollar reforzamos y aprendimos lo visto en clase, dos dimos cuenta de lo útil que puede llegar a ser los DataFrames y BigData si se utiliza correctamente.