**Pasos para conectar la dataset descargada**

Se crea una carpeta en el sistema HDFS utilizando el comando

hdfs dfs -mkdir /AmazonAWS

wget "https://storage.googleapis.com/kaggle-data-sets/1985/3404/compressed/data.csv.zip?X-Goog-Algorithm=GOOG4-RSA-SHA256&X-Goog-Credential=gcp-kaggle-com%40kaggle-161607.iam.gserviceaccount.com%2F20251010%2Fauto%2Fstorage%2Fgoog4\_request&X-Goog-Date=20251010T034713Z&X-Goog-Expires=259200&X-Goog-SignedHeaders=host&X-Goog-Signature=" -O data.csv.zip

unzip data.csv.zip

Ahora copiamos el archivo del Dataset descargado a la carpeta HDFS que creamos

hdfs dfs -put /home/hadoop/data.csv /AmazonAWS

Podemos también por línea de comando ver el archivo del dataset en la carpeta HDFS llamada AmazonAWS

hdfs dfs -ls /AmazonAWS

Se crea un archivo de python con extension .py lo podemos llamar AmazonAWS.py, donde colocaremos el código del programa que se conectará al dataset descargado dentro del sistema HDFS, luego convertiremos estos datos en un dataframe de Spark y sobre este dataframe se realizarán diferentes operaciones y consultas.

**nano AmazonAWS.py**

Pegamos el siguiente código:

**# ============================================================**

**# Procesamiento de datos de Amazon AWS con Apache Spark (Batch)**

**# ============================================================**

**# Importar librerías necesarias**

**from pyspark.sql import SparkSession**

**from pyspark.sql import functions as F**

**# Inicializar la sesión de Spark**

**spark = SparkSession.builder \**

**.appName("AmazonAWS") \**

**.getOrCreate()**

**# Definir la ruta del archivo CSV en HDFS**

**file\_path = "hdfs://localhost:9000/AmazonAWS/data.csv"**

**# Leer el archivo CSV desde HDFS**

**df = spark.read.format("csv") \**

**.option("header", "true") \**

**.option("inferSchema", "true") \**

**.load(file\_path)**

**# Imprimir el esquema del DataFrame**

**print(" Esquema del DataFrame:")**

**df.printSchema()**

**# Mostrar las primeras filas del DataFrame**

**print(" Primeras filas del conjunto de datos:")**

**df.show(10, truncate=False)**

**# Mostrar estadísticas básicas de las columnas numéricas**

**print(" Estadísticas básicas del conjunto de datos:")**

**df.summary().show()**

**# Finalizar la sesión**

**spark.stop()**

dar Crtl+O enter y luego Crtl+X para salir

para ejecutar el archivo ingresamos la siguiente instrucción:

start-master.sh

$HADOOP\_HOME/sbin/start-all.sh

python3 Amaz