Tutorial for uploading to Hadoop

Step 1: Uploading the Dataset to Hadoop

- **Purpose**: To download the Chicago Crime dataset from Kaggle and upload it to the Hadoop file system (HDFS).
- **Guide**:
    - Create a Kaggle account if you don't have one already.
    - Download the dataset from Kaggle using the link: https://www.kaggle.com/datasets/n3v375/chicago-crime-from-01jan2001-to-22jul2020
    - Unzip the downloaded archive file.
    - Locate the CSV file named "Crimes_-_2001_to_Present.csv".
    - Move this CSV file to your Hadoop user directory.

Step 2: Securely Copying Files to Hadoop Cluster

- **Purpose**: To securely copy files between your local machine and the Hadoop cluster.
- **Guide**:
    - Use the **scp** command in a terminal window (Git Bash): scp /Users/Francisco/Crimes_-_2001_to_Present.csv ffigue10@129.146.148.35:/home/ffigue10/
    - Replace placeholders (username and IP address) with your actual Hadoop server details.
    - Ensure you have the necessary permissions to access the destination directory.

Step 3: Connect to Hadoop Cluster and Upload Dataset to HDFS

- **Purpose**: To connect to the Hadoop cluster and upload the dataset to HDFS.
- **Guide**:
    - Open a terminal and SSH into the Hadoop cluster: ssh yourusername@ip address
    - Create a directory in HDFS using the **hdfs dfs -mkdir** command: hdfs dfs -mkdir Crimes
    - Use **hdfs dfs -put** to upload the CSV file to the created directory in HDFS: hdfs dfs -put Crimes_-_2001_to_Present.csv Crimes/
    - Confirm that the file is successfully uploaded using **hdfs dfs -ls**: hdfs dfs -ls Crimes/

Step 4: Open Beeline

- **Purpose**: To open Beeline for executing Hive queries.
- **Guide**:
    - Open a new terminal.
    - Type **beeline** and press Enter.

Step 5: Run show tables;

- **Purpose**: To check existing tables in Hive.
- **Guide**:
    - In the Beeline terminal, type **show tables;** and press Enter.

Step 6: Run DROP TABLE IF EXISTS crime;

- **Purpose**: To drop the "crime" table if it already exists in Hive.
- **Guide**:
    - In the Beeline terminal, type **DROP TABLE IF EXISTS crime;** and press Enter.

Step 7: Run CREATE EXTERNAL TABLE

- **Purpose**: To create an external table in Hive for storing the Chicago Crime dataset.
- **Guide**:
    - Execute the provide Hive query to create the table: CREATE EXTERNAL TABLE IF NOT EXISTS crime(ID INT, Case_Number STRING, Date_Time STRING, Block STRING, IUCR STRING, Primary_Type STRING, Description STRING, Location_Description STRING, Arrest BOOLEAN, Domestic BOOLEAN, Beat INT, District INT, Ward INT, Community_Area INT, FBI_Code STRING, X_Coordinate INT, Y_Coordinate INT, Year STRING, Updated_On TIMESTAMP, Latitude DOUBLE, Longitude DOUBLE, Location STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION '/user/mtoyinb/Crimes' TBLPROPERTIES ('skip.header.line.count'='1');
    - Replace placeholders with appropriate details if needed.

Step 8: Run SELECT * from crime limit 5;

- **Purpose**: To check if the "crime" table is created successfully by displaying the first 5 rows.
- **Guide**:
    - In the Beeline terminal, type **SELECT * from crime limit 5;** and press Enter.