ScrapeStockMarketData

*Install Necessary libraries*

# pip install bs4

# pip install bs4

# pip install requests

# pip install pandas

from bs4 import BeautifulSoup

import requests

import pandas as pd

import numpy as np

from pyspark.sql.functions import col, current\_date, lit

from pyspark.sql.types import StringType, StructType, StructField, DoubleType, DecimalType, FloatType

HTTP Header is required - User Agent help to identify the user to access the content (https://www.whatismybrowser.com/detect/what-is-my-user-agent/)

dbutils.widgets.dropdown("TIMELINE", "weekly", ["weekly", "daily", "monthly"])

v\_timeline = dbutils.widgets.get("TIMELINE")

dbutils.widgets.dropdown("NSE\_BSE", "nse", ["nse", "bsc"])

v\_nse\_bse = dbutils.widgets.get("NSE\_BSE")

dbutils.widgets.dropdown("category", "groupall", ["nifty", "groupall"])

v\_category = dbutils.widgets.get("category")

v\_comments = v\_timeline + " " + v\_nse\_bse + " " + v\_category

URL = f"https://money.rediff.com/gainers/{v\_nse\_bse}/{v\_timeline}/{v\_category}"

#To display

URL

# add your user agent

HEADERS = ({'User-Agent':'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/107.0.0.0 Safari/537.36', 'Accept-Language': 'en-US, en;q=0.5'})

# HTTP Request

webpage = requests.get(URL, headers=HEADERS)

# Response must be 200

webpage

Out[6]: <Response [200]>

soup = BeautifulSoup(webpage.content, "html.parser")

# Creating a dictionary

d = {"company":[], "Prev\_Close":[], "Current\_Price":[], "Changes":[]}

# Fetch links as List of Tag Objects

links\_tr = soup.find\_all("tr")

for i in links\_tr[1:-2]:

links\_td = i.find\_all("td" )

links\_values = links\_td[0].find\_all("a")

var1 = links\_values[0].get\_text()

var2 = links\_td[1].get\_text()

var3 = links\_td[2].get\_text()

var4 = links\_td[3].get\_text()

d['company'].append(var1[6:-6].strip())

d['Prev\_Close'].append(var2.strip() )

d['Current\_Price'].append(var3.strip() )

d['Changes'].append(float(var4[1:].strip()) )

#print(float(var4[1:].strip()))

stock\_df = pd.DataFrame.from\_dict(d)

data=[["1"]]

df=spark.createDataFrame(data,["id"])

df1 = df.withColumn("current\_date",current\_date())

cur\_date = df1.collect()

v\_cur\_date = (cur\_date[0].\_\_getitem\_\_('current\_date'))

v\_comments = v\_comments + " " + str(v\_cur\_date)

stockdata\_schema = StructType(fields = [

StructField("company", StringType(), False),

StructField("Prev\_Close", StringType(), False),

StructField("Current\_Price", StringType(), False),

StructField("Changes", FloatType(), False)

])

# Creating a new Data Frame from dictionary

stdf = spark.createDataFrame(stock\_df, stockdata\_schema)

stdf\_sorted = stdf.sort(col("Changes").desc()).withColumn("comments", lit(v\_comments))

stdf\_sorted.show()

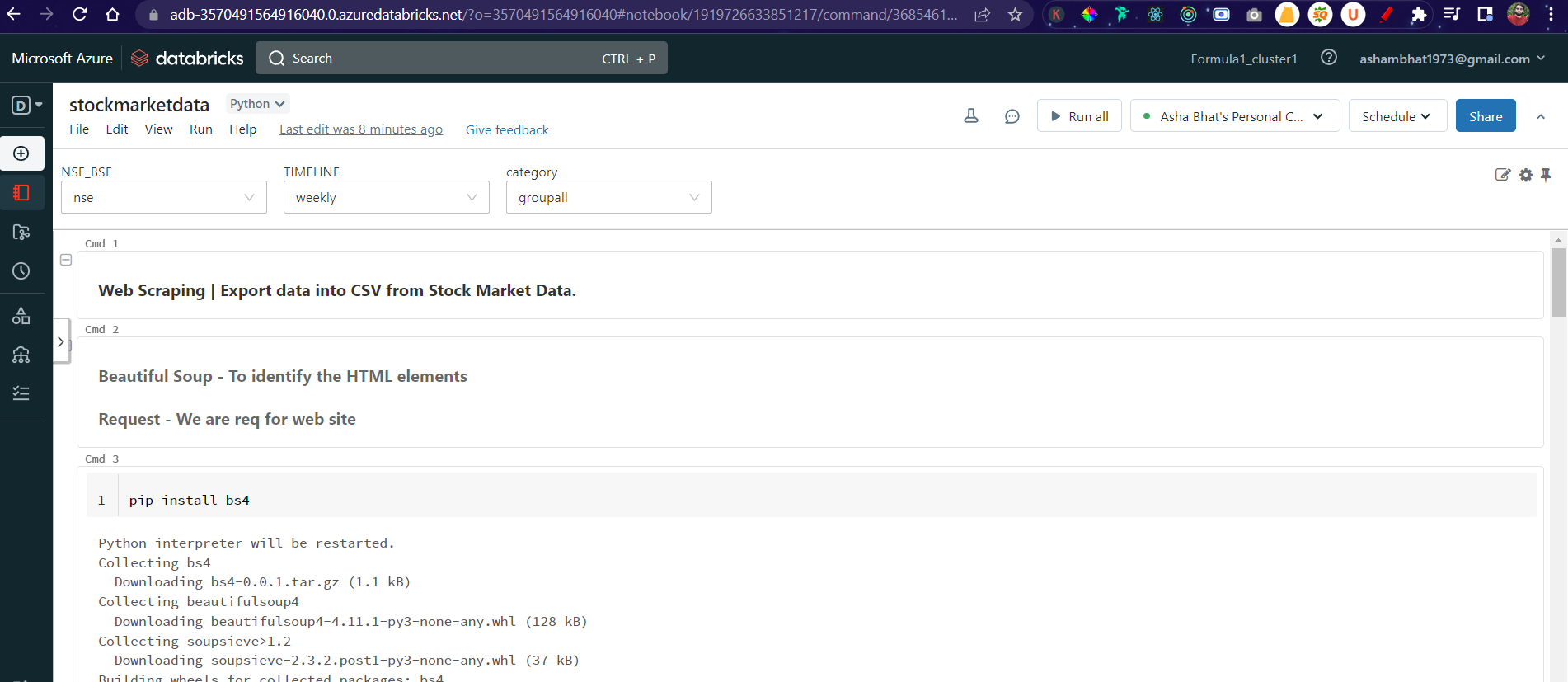
stdf\_sorted.write\

.mode("overwrite")\

.option("header",True)\

.csv(f'/mnt/formula1datalake97/stockmarket/{v\_cur\_date}.csv')

Add 3 Input Text Box



The data exported as CSV

