gold_local

December 23, 2020

1 Làm sạch dữ liệu và đưa dữ liệu lên HDFS

[1]:
```
from pyspark.context import SparkContext
from pyspark.sql.session import SparkSession
```

[2]:
```
#Truy cập vào master của cụm tại local với tên chương trình là Gold0
spark = SparkSession.builder.master("local").appName("Gold0").getOrCreate()
```

[3]:
```
#Đọc file newPrices.csv tại local
csvFile = spark.read.format("csv")\
.option("header", "true")\
.option("inferSchema", "true")\
.load("./newPrices.csv")
```

[4]:
```
#Lấy thông tin Schema
csvFile.printSchema()
```

```
root
 |-- Date: string (nullable = true)
 |-- US_dollar: string (nullable = true)
 |-- Euro: string (nullable = true)
 |-- Japanese_yen: string (nullable = true)
 |-- Vietnamese_dong: string (nullable = true)
 |-- Korean_won: string (nullable = true)
```

[5]:
```
#Trả về số record
csvFile.count()
```

[5]:
```
10941
```

[6]:
```
#In ra record
csvFile.show()
```

```
+----------+---------+-----+------------+---------------+----------+
|      Date|US_dollar| Euro|Japanese_yen|Vietnamese_dong|Korean_won|
+----------+---------+-----+------------+---------------+----------+
|12/29/1978|      226|137.1|        #N/A|           #N/A|      #N/A|
```

```
|  1/1/1979|      226|137.1|        #N/A|           #N/A|      #N/A|
|  1/2/1979|    226.8|137.3|   43,164.90|           #N/A|107,470.00|
|  1/3/1979|    218.6|  134|   43,717.90|           #N/A|108,027.40|
|  1/4/1979|    223.2|136.8|   43,674.90|           #N/A|108,602.50|
|  1/5/1979|    225.5|138.4|   44,582.50|           #N/A|110,510.40|
|  1/8/1979|    223.1|136.4|   44,436.20|           #N/A|110,356.30|
|  1/9/1979|      224|137.3|   44,045.60|           #N/A|109,248.40|
| 1/10/1979|    220.7|135.5|   43,366.40|           #N/A|108,108.30|
| 1/11/1979|    220.7|135.9|   43,770.60|           #N/A|108,771.70|
| 1/12/1979|    217.6|134.1|   42,837.10|           #N/A|106,856.60|
| 1/15/1979|    216.9|133.8|   42,795.30|           #N/A|106,819.80|
| 1/16/1979|    220.7|135.6|   43,225.90|           #N/A|107,689.80|
| 1/17/1979|    227.3|139.2|   44,349.70|           #N/A|110,419.80|
| 1/18/1979|    231.8|141.7|   44,823.40|           #N/A|111,599.20|
| 1/19/1979|    230.6|141.1|   46,908.80|           #N/A|116,335.50|
| 1/22/1979|      235|  144|   46,387.40|           #N/A|115,493.30|
| 1/23/1979|      230|141.1|   45,633.10|           #N/A|113,615.30|
| 1/24/1979|    236.1|144.7|   46,372.40|           #N/A|115,456.00|
| 1/25/1979|    233.9|  144|   46,890.30|           #N/A|116,157.90|
+----------+---------+-----+------------+---------------+----------+
only showing top 20 rows
```

[7]:
```
#Lấy ra list tên các cột
cols=csvFile.columns
```

Ta cần thay thế các giá trị string "#N/A" bằng các giá trị "0" và loại bỏ kí tự "," giữa các giá trị để chuẩn bị cho việc ép kiểu sang float cho các cột dữ liệu

[8]:
```
from pyspark.sql.functions import *
from pyspark.sql.types import FloatType

index =1
while (index<len(cols)):
    csvFile = csvFile.withColumn(cols[index], regexp_replace(cols[index], '#N/
 ↪A', '0'))
    csvFile = csvFile.withColumn(cols[index], regexp_replace(cols[index], ',',␣
 ↪''))
    csvFile = csvFile.withColumn(cols[index], csvFile[cols[index]].
 ↪cast('float'))
    index=index+1
```

[9]:
```
csvFile.printSchema()
```

```
root
 |-- Date: string (nullable = true)
 |-- US_dollar: float (nullable = true)
 |-- Euro: float (nullable = true)
```

```
 |-- Japanese_yen: float (nullable = true)
 |-- Vietnamese_dong: float (nullable = true)
 |-- Korean_won: float (nullable = true)
```

Chuyển đổi kiểu dữ liệu của cột "Date" từ String sang Date

[10]:
```
from datetime import datetime
from pyspark.sql.functions import col, udf
from pyspark.sql.types import DateType

#Convert dữ liệu về form của kiểu date:
func =  udf (lambda x: datetime.strptime(x, '%m/%d/%Y'), DateType())

csvFile = csvFile.withColumn('Date', func(col('Date')))

csvFile.printSchema()
```

```
root
 |-- Date: date (nullable = true)
 |-- US_dollar: float (nullable = true)
 |-- Euro: float (nullable = true)
 |-- Japanese_yen: float (nullable = true)
 |-- Vietnamese_dong: float (nullable = true)
 |-- Korean_won: float (nullable = true)
```

[11]:
```
csvFile.show()
```

```
+----------+---------+-----+------------+---------------+----------+
|      Date|US_dollar| Euro|Japanese_yen|Vietnamese_dong|Korean_won|
+----------+---------+-----+------------+---------------+----------+
|1978-12-29|    226.0|137.1|         0.0|            0.0|       0.0|
|1979-01-01|    226.0|137.1|         0.0|            0.0|       0.0|
|1979-01-02|    226.8|137.3|     43164.9|            0.0|  107470.0|
|1979-01-03|    218.6|134.0|     43717.9|            0.0|  108027.4|
|1979-01-04|    223.2|136.8|     43674.9|            0.0|  108602.5|
|1979-01-05|    225.5|138.4|     44582.5|            0.0|  110510.4|
|1979-01-08|    223.1|136.4|     44436.2|            0.0|  110356.3|
|1979-01-09|    224.0|137.3|     44045.6|            0.0|  109248.4|
|1979-01-10|    220.7|135.5|     43366.4|            0.0|  108108.3|
|1979-01-11|    220.7|135.9|     43770.6|            0.0|  108771.7|
|1979-01-12|    217.6|134.1|     42837.1|            0.0|  106856.6|
|1979-01-15|    216.9|133.8|     42795.3|            0.0|  106819.8|
|1979-01-16|    220.7|135.6|     43225.9|            0.0|  107689.8|
|1979-01-17|    227.3|139.2|     44349.7|            0.0|  110419.8|
|1979-01-18|    231.8|141.7|     44823.4|            0.0|  111599.2|
|1979-01-19|    230.6|141.1|     46908.8|            0.0|  116335.5|
|1979-01-22|    235.0|144.0|     46387.4|            0.0|  115493.3|
```

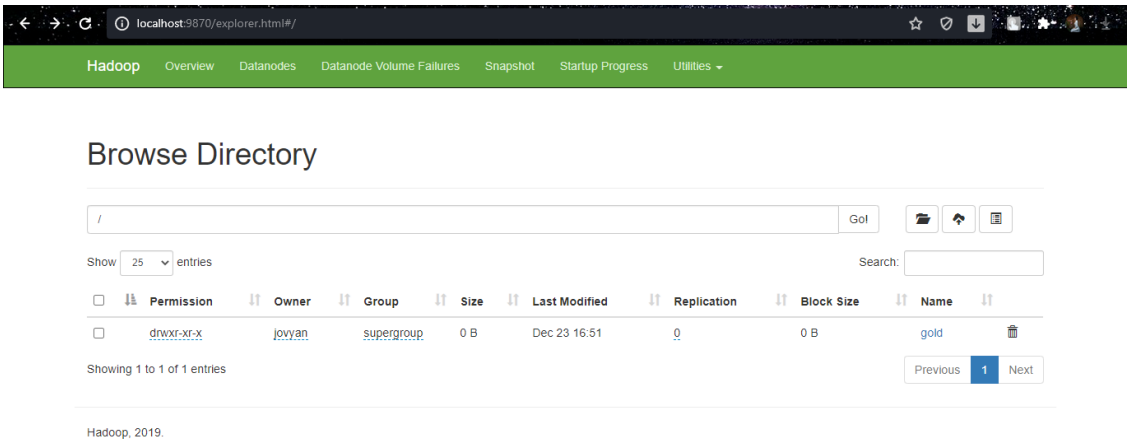
```
|1979-01-23|    230.0|141.1|     45633.1|            0.0|  113615.3|
|1979-01-24|    236.1|144.7|     46372.4|            0.0|  115456.0|
|1979-01-25|    233.9|144.0|     46890.3|            0.0|  116157.9|
+----------+---------+-----+------------+---------------+----------+
only showing top 20 rows
```

[12]:
```
#Ghi vào file gold.csv trên hdfs
csvFile.write.format("csv").mode("overwrite").option("header", "true")\
.save("hdfs://namenode:9000/gold/gold.csv")
```

Truy cập giao diện Web cho namenode tại http://localhost:9870 ta sẽ thấy file gold.csv được load lên ở trong folder gold

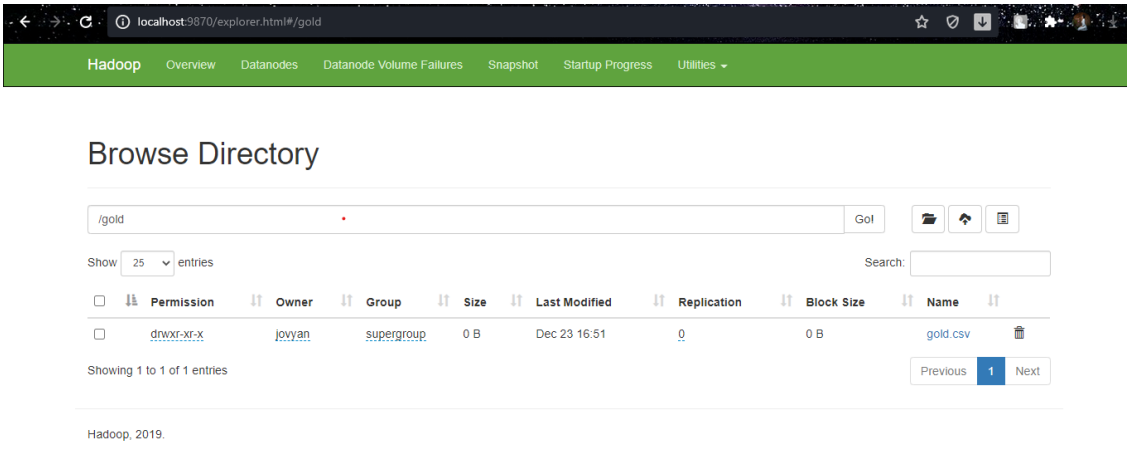
[13]:
```
from IPython.display import Image
Image("./image/gold_folder.png")
```

[13]:



[14]:
```
Image("./image/file_gold.png")
```

[14]:


localhost:9870/explorer.html#/gold
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/gold
Go!
Show 25 entries
Search:
Permission
Owner
Group
Size
Last Modified
Replication
Block Size
Name
drwxr-xr-x
jovyan
supergroup
0 B
Dec 23 16:51
0
0 B
gold.csv
Showing 1 to 1 of 1 entries
Previous
1
Next
Hadoop, 2019.

[ ]: