Flatten Json to Tabluar using in Pyspark

Created by	vicky waran
Created time	@January 4, 2024 4:14 PM
Status	Done
Tags	pyspark youtube

Project Summary

Python Name :Flatten Json to tabluar using in pyspark

Referral Youtube : **https://www.youtube.com/watch?v=FT0MQNBaoqo&t=8s**

Location in local :C:\Users\admin\OneDrive\Documents\Python\py and pyspark\Jupyter\Project_flaten_json

Github:

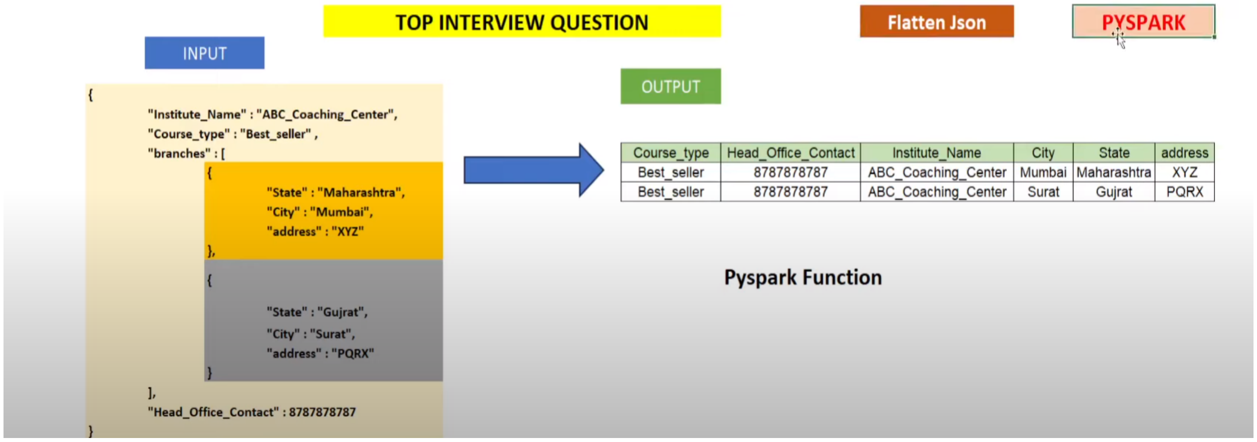
FORMULA:

#FORMULA : WHENEVER WE SEE A ARRAY USE THE EXPLODE , STRUCT USE COLUMN_NAME.*

ARRAY —> explode(), for struct —→ column.*

Goals

We will be Converting the Json Fomrat file Details to Tabular Fomrat using Pyspark functions



TOP INTERVIEW QUESTION

Flatten Json

PYSPARK

INPUT

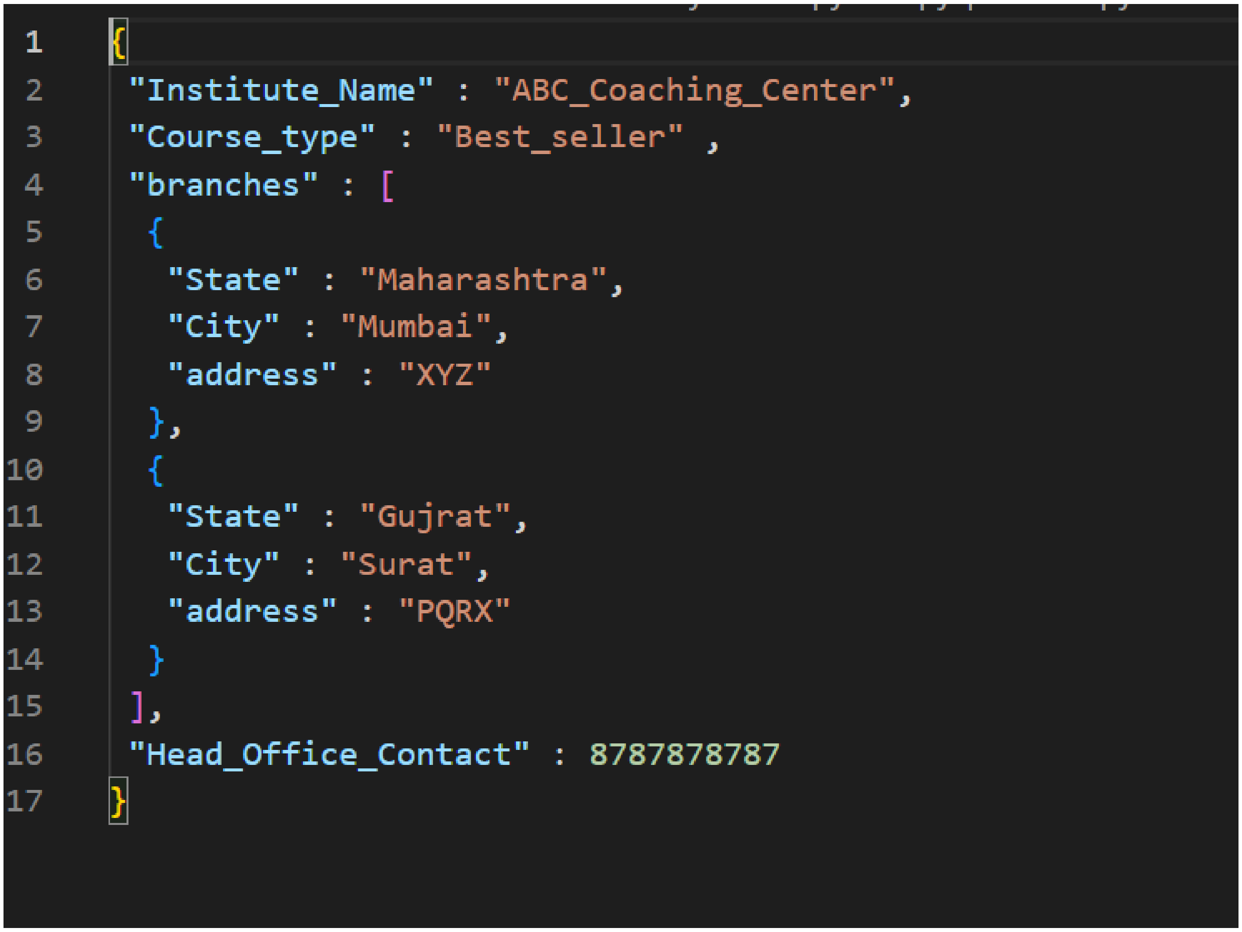
```
{
        "Institute_Name" : "ABC_Coaching_Center",
        "Course_type" : "Best_seller" ,
        "branches" : [
                {
                        "State" : "Maharashtra",
                        "City" : "Mumbai",
                        "address" : "XYZ"
                },
                {
                        "State" : "Gujrat",
                        "City" : "Surat",
                        "address" : "PQRX"
                }
        ],
        "Head_Office_Contact" : 8787878787
}
```

OUTPUT

Course_type	Head_Office_Contact	Institute_Name	City	State	address
Best_seller	8787878787	ABC_Coaching_Center	Mumbai	Maharashtra	XYZ
Best_seller	8787878787	ABC_Coaching_Center	Surat	Gujrat	PQRX

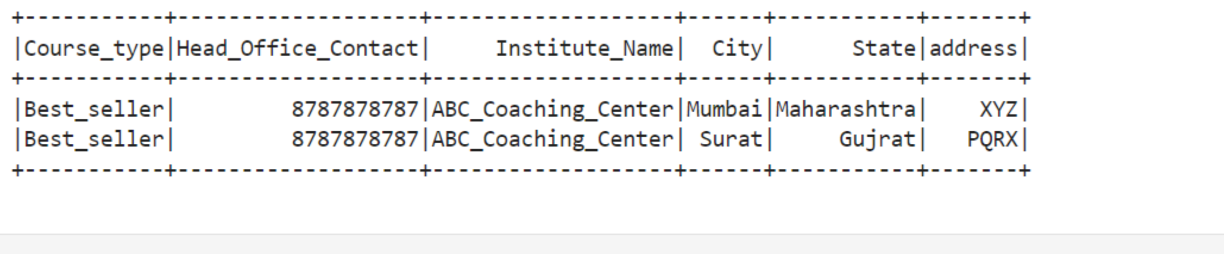
Pyspark Function

Input:



```
{
 "Institute_Name" : "ABC_Coaching_Center",
 "Course_type" : "Best_seller" ,
 "branches" : [
  {
   "State" : "Maharashtra",
   "City" : "Mumbai",
   "address" : "XYZ"
  },
  {
   "State" : "Gujrat",
   "City" : "Surat",
   "address" : "PQRX"
  }
 ],
 "Head_Office_Contact" : 8787878787
}
```

Output:



```
+-----------+-------------------+-------------------+------+-----------+-------+
|Course_type|Head_Office_Contact|     Institute_Name|  City|      State|address|
+-----------+-------------------+-------------------+------+-----------+-------+
|Best_seller|         8787878787|ABC_Coaching_Center|Mumbai|Maharashtra|    XYZ|
|Best_seller|         8787878787|ABC_Coaching_Center| Surat|     Gujrat|   PQRX|
+-----------+-------------------+-------------------+------+-----------+-------+
```

Function We used:

pyspark:

explode() - it takes the array input and converts into the sequences format without any for loop.

df_json=spark.read.format("json").option("multiline",True).load("sample.json") #adding option multiline since multiline json

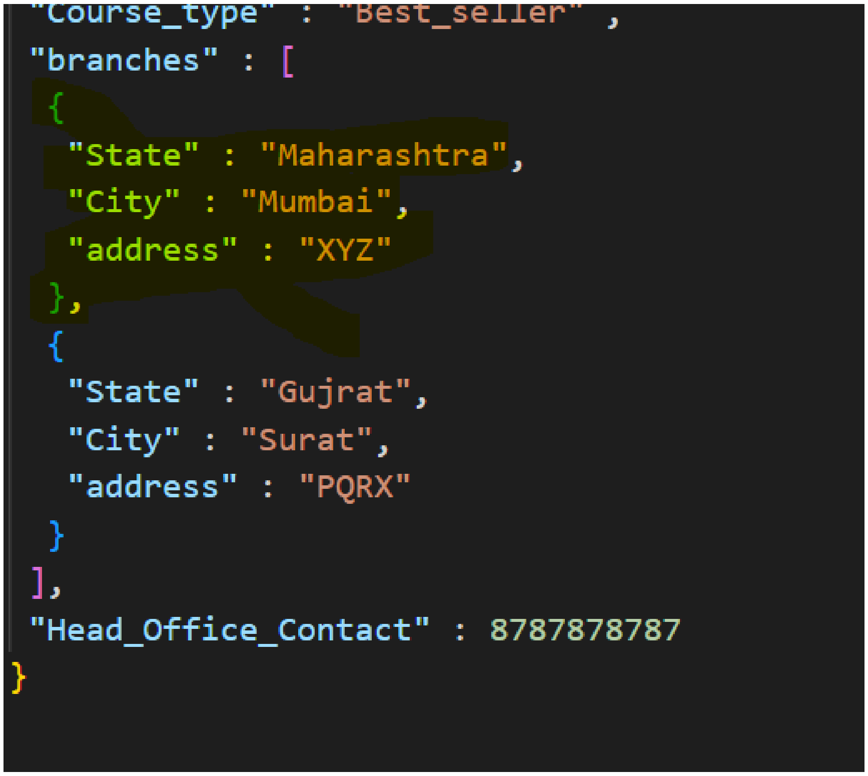
df_json.printSchema() - it prints the schema of the dataframe

from pyspark.sql.functions import explode,col - for importing the explode and col function

df_exploded =df_json.select("Course_type","Head_Office_Contact","Institute_Name",explode(col("branches"))) - to explod the branches in the two separate Section

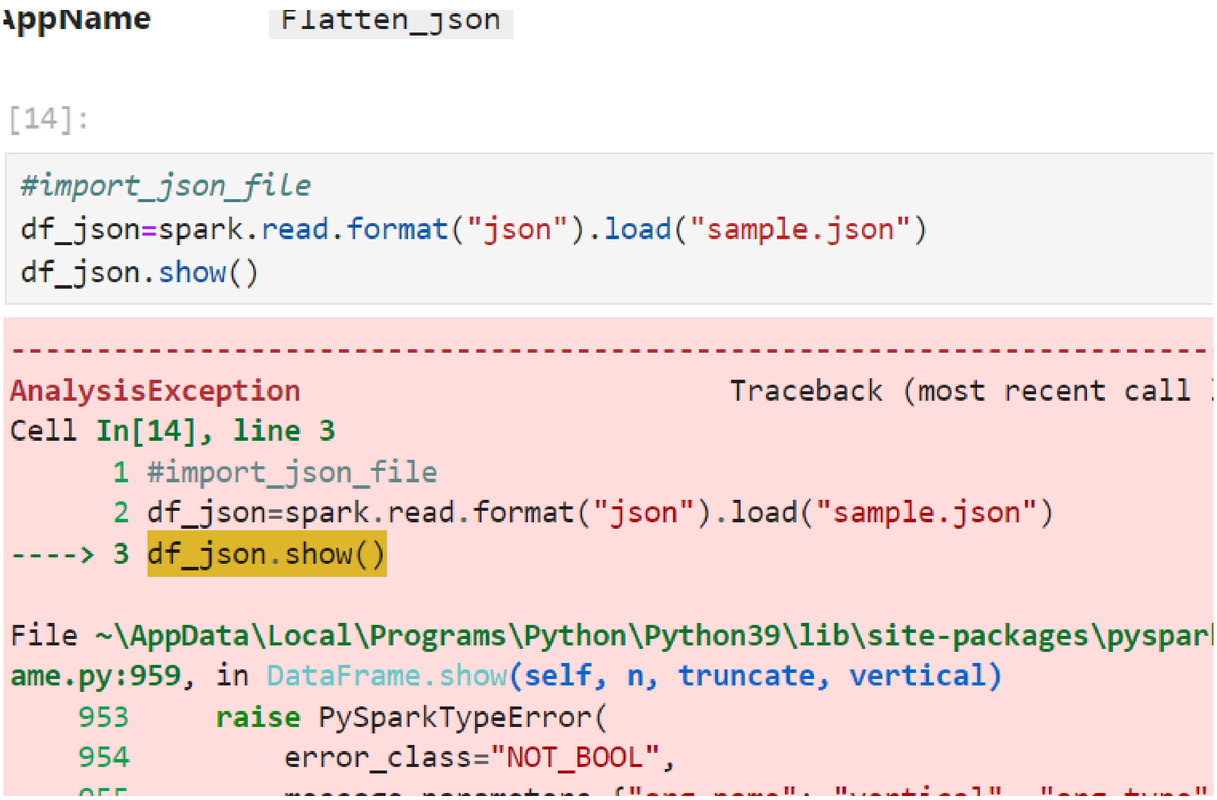
Things we learned while Coding:

Inside this Branches we call it as a struct



```
 "Course_type" : "Best_seller" ,
 "branches" : [
  {
   "State" : "Maharashtra",
   "City" : "Mumbai",
   "address" : "XYZ"
  },
  {
   "State" : "Gujrat",
   "City" : "Surat",
   "address" : "PQRX"
  }
 ],
 "Head_Office_Contact" : 8787878787
}
```

Input file is multiline Json since we are getting a error



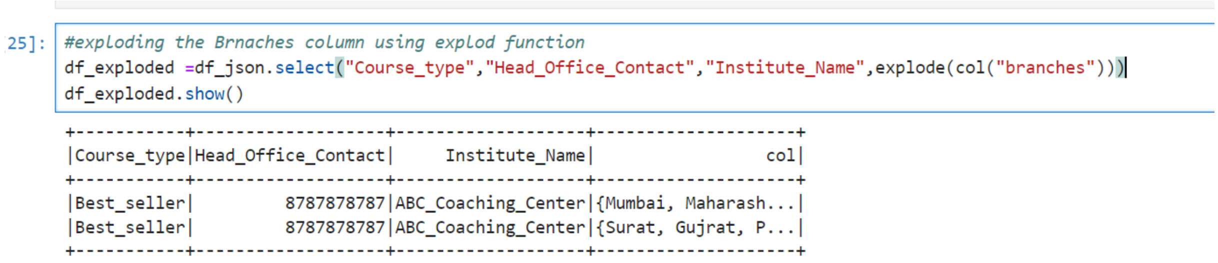
```
AppName        Flatten_json

[14]:
#import_json_file
df_json=spark.read.format("json").load("sample.json")
df_json.show()
```

```
---------------------------------------------------------------------------
AnalysisException                         Traceback (most recent call
Cell In[14], line 3
      1 #import_json_file
      2 df_json=spark.read.format("json").load("sample.json")
----> 3 df_json.show()

File ~\AppData\Local\Programs\Python\Python39\lib\site-packages\pyspar
ame.py:959, in DataFrame.show(self, n, truncate, vertical)
    953     raise PySparkTypeError(
    954         error_class="NOT_BOOL",
    955         message_parameters={"arg_name": "vertical", "arg_type"
```

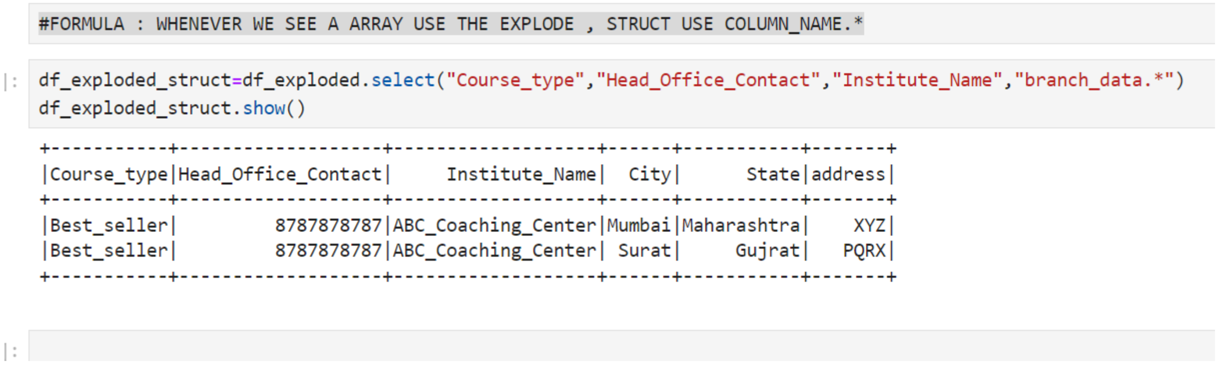
To explode the branches in the two rows use explode function with col



```
25]: #exploding the Brnaches column using explod function
     df_exploded =df_json.select("Course_type","Head_Office_Contact","Institute_Name",explode(col("branches")))
     df_exploded.show()

     +-----------+-------------------+------------------+--------------------+
     |Course_type|Head_Office_Contact|    Institute_Name|                 col|
     +-----------+-------------------+------------------+--------------------+
     |Best_seller|         8787878787|ABC_Coaching_Center|{Mumbai, Maharash...|
     |Best_seller|         8787878787|ABC_Coaching_Center|{Surat, Gujrat, P...|
     +-----------+-------------------+------------------+--------------------+
```

use column.* in the explode to make the key and value visible in the table



```
#FORMULA : WHENEVER WE SEE A ARRAY USE THE EXPLODE , STRUCT USE COLUMN_NAME.*

 |: df_exploded_struct=df_exploded.select("Course_type","Head_Office_Contact","Institute_Name","branch_data.*")
    df_exploded_struct.show()

    +-----------+-------------------+-------------------+------+-----------+-------+
    |Course_type|Head_Office_Contact|     Institute_Name|  City|      State|address|
    +-----------+-------------------+-------------------+------+-----------+-------+
    |Best_seller|         8787878787|ABC_Coaching_Center|Mumbai|Maharashtra|    XYZ|
    |Best_seller|         8787878787|ABC_Coaching_Center| Surat|     Gujrat|   PQRX|
    +-----------+-------------------+-------------------+------+-----------+-------+

 |:
```