In [4]:
```
from pyspark.sql.functions import *
from pyspark.sql.types import *
```

In [8]:
```
from pyspark import SparkConf, SparkContext
```

In [9]:
```
from pyspark.sql import SparkSession
from pyspark.sql import Row
from pyspark.sql import functions as func
```

In [10]:
```
spark = SparkSession.builder.appName("ratings").getOrCreate()
```

In [5]:
```
def parse_nested_json(df):
  col_lst = []
  print('Parsing the dataframe....')
  for cl in df.schema.names:
    aa= df.schema[cl].dataType
    if isinstance(aa, ArrayType):
      print("Inside tghe col name {0} of type 'ArrayType' ".format(cl))
      df = df.withColumn(cl, explode(cl).alias(cl))
      col_lst.append(cl)

    elif isinstance(df.schema[cl].dataType, StructType):
      print(("Inside tghe col name {0} of type 'ArrayType'".format(cl)))
      for fld in df.schema[cl].dataType.fields:
        col_lst.append(col(cl+'.'+fld.name).alias(cl+'_'+fld.name))
    else:
      col_lst.append(cl)
  print('____''25)
  print(col_lst)
  df = df.select(col_lst)
  return df
```

In [11]:
```
df = spark.read.option("multiline","true").json('correct.json')
```

In [12]:
```
read_nested_json_flag = True
```

In [13]:
```
while read_nested_json_flag:
  print("Reading Nested JSON File ... ")
  df = parse_nested_json(df)
  read_nested_json_flag = False

  if df.count() > 1:
    df = df

  for column_name in df.schema.names:
    if isinstance(df.schema[column_name].dataType, ArrayType):
      read_nested_json_flag = True
    elif isinstance(df.schema[column_name].dataType, StructType):
      read_nested_json_flag = True
```

```
Reading Nested JSON File ...
Parsing the dataframe....
Inside tghe col name children of type 'ArrayType'

['EventTS', 'EventTypeCd', 'FormCd', 'children', 'guid', 'industry', 'is_service_provider',
'is_subscribed', 'name', 'rating', 'rating_type']
Reading Nested JSON File ...
Parsing the dataframe....
Inside tghe col name children of type 'ArrayType'

['EventTS', 'EventTypeCd', 'FormCd', Column<'children.children AS `children_children`'>, Colu
mn<'children.guid AS `children_guid`'>, Column<'children.industry AS `children_industry`'>, C
olumn<'children.is_service_provider AS `children_is_service_provider`'>, Column<'children.is_
subscribed AS `children_is_subscribed`'>, Column<'children.name AS `children_name`'>, Column
<'children.rating AS `children_rating`'>, Column<'children.rating_type AS `children_rating_ty
pe`'>, 'guid', 'industry', 'is_service_provider', 'is_subscribed', 'name', 'rating', 'rating_
type']
Reading Nested JSON File ...
Parsing the dataframe....
Inside tghe col name children_children of type 'ArrayType'
____________________________________________________________________________________________

['EventTS', 'EventTypeCd', 'FormCd', 'children_children', 'children_guid', 'children_industr
y', 'children_is_service_provider', 'children_is_subscribed', 'children_name', 'children_rati
ng', 'children_rating_type', 'guid', 'industry', 'is_service_provider', 'is_subscribed', 'nam
e', 'rating', 'rating_type']
Reading Nested JSON File ...
Parsing the dataframe....
Inside tghe col name children_children of type 'ArrayType'
____________________________________________________________________________________________

['EventTS', 'EventTypeCd', 'FormCd', Column<'children_children_children AS `children_children
_children`'>, Column<'children_children_guid AS `children_children_guid`'>, Column<'children_
children_industry AS `children_children_industry`'>, Column<'children_children.is_service_pro
vider AS `children_children_is_service_provider`'>, Column<'children_children_is_subscribed A
S `children_children_is_subscribed`'>, Column<'children_children_name AS `children_children_n
ame`'>, Column<'children_children_rating AS `children_children_rating`'>, Column<'children_ch
ildren_rating_type AS `children_children_rating_type`'>, 'children_guid', 'children_industr
y', 'children_is_service_provider', 'children_is_subscribed', 'children_name', 'children_rati
ng', 'children_rating_type', 'guid', 'industry', 'is_service_provider', 'is_subscribed', 'nam
e', 'rating', 'rating_type']
Reading Nested JSON File ...
Parsing the dataframe....
Inside tghe col name children_children_children of type 'ArrayType'

['EventTS', 'EventTypeCd', 'FormCd', 'children_children_children', 'children_children_guid',
'children_children_industry', 'children_children_is_service_provider', 'children_children_is_
subscribed', 'children_children_name', 'children_children_rating', 'children_children_rating_
type', 'children_guid', 'children_industry', 'children_is_service_provider', 'children_is_sub
scribed', 'children_name', 'children_rating', 'children_rating_type', 'guid', 'industry', 'is
_service_provider', 'is_subscribed', 'name', 'rating', 'rating_type']
Reading Nested JSON File ...
Parsing the dataframe....
Inside tghe col name children_children_children of type 'ArrayType'

['EventTS', 'EventTypeCd', 'FormCd', Column<'children_children_children_children AS `children
_children_children_children`'>, Column<'children_children_children_guid AS `children_children
_children_guid`'>, Column<'children_children_children_industry AS `children_children_children
_industry`'>, Column<'children_children_children_is_service_provider AS `children_children_ch
ildren_is_service_provider`'>, Column<'children_children_children_is_subscribed AS `children_
children_children_is_subscribed`'>, Column<'children_children_children_name AS `children_chil
dren_children_name`'>, Column<'children_children_children_rating AS `children_children_childr
en_rating`'>, Column<'children_children_children_rating_type AS `children_children_children_r
ating_type`'>, 'children_children_guid', 'children_children_industry', 'children_children_is_
service_provider', 'children_children_is_subscribed', 'children_children_name', 'children_chi
ldren_rating', 'children_children_rating_type', 'children_guid', 'children_industry', 'childr
en_is_service_provider', 'children_is_subscribed', 'children_name', 'children_rating', 'child
ren_rating_type', 'guid', 'industry', 'is_service_provider', 'is_subscribed', 'name', 'ratin
g', 'rating_type']
Reading Nested JSON File ...
Parsing the dataframe....
Inside tghe col name children_children_children_children of type 'ArrayType'
____________________________________________________________________________________________

['EventTS', 'EventTypeCd', 'FormCd', 'children_children_children_children', 'children_childre
n_children_guid', 'children_children_children_industry', 'children_children_children_is_servi
ce_provider', 'children_children_children_is_subscribed', 'children_children_children_name',
'children_children_children_rating', 'children_children_children_rating_type', 'children_chil
dren_guid', 'children_children_industry', 'children_children_is_service_provider', 'children_
children_is_subscribed', 'children_children_name', 'children_children_rating', 'children_chil
dren_rating_type', 'children_guid', 'children_industry', 'children_is_service_provider', 'chi
ldren_is_subscribed', 'children_name', 'children_rating', 'children_rating_type', 'guid', 'in
dustry', 'is_service_provider', 'is_subscribed', 'name', 'rating', 'rating_type']
```

In [14]:
```
## registering  the dataframe as temp table and creating the normalized table to use sql
df1.registerTempTable('aa')
```

In [21]:
```
spark.sql("""
select
null as parent
, null as parent_name
, guid
, industry
, is_service_provider
, is_subscribed
, name
, rating
, rating_type
, EventTS
, EventTypeCd
, FormCd
from aa

union all

select
children_children_guid as parent
, children_children_name as parent_name
, children_children_children_guid as guid
, children_children_children_industry as industry
, children_children_children_is_service_provider as is_service_provider
, children_children_children_is_subscribed  as is_subscribed
, children_children_children_name as name
, children_children_children_rating as rating
, children_children_children_rating_type rating_type
, EventTS
, EventTypeCd
, FormCd
from aa

union all

select
children_guid as parent
, children_name as parent_name
, children_children_guid as guid
, children_children_industry as industry
, children_children_is_service_provider as is_service_provider
, children_children_is_subscribed  as is_subscribed
, children_children_name as name
, children_children_rating as rating
, children_children_rating_type rating_type
, EventTS
, EventTypeCd
, FormCd
from aa

union all

select
guid as parent
,name as parent_name
, children_guid as guid
, children_industry as industry
, children_is_service_provider as is_service_provider
, children_is_subscribed  as is_subscribed
, children_name as name
, children_rating as rating
, children_rating_type rating_type
, EventTS
, EventTypeCd
, FormCd
from aa
""").registerTempTable('prsd_json')
```

In [37]:
```
data = spark.sql('select * from prsd_json').coalesce(1).toPandas()
```

In [40]:
```
data.to_csv('file.csv')# parsing file to csv
```

In [25]:
```
q1_df = spark.sql('select industry, max(rating) as max_rating from prsd_json group  by indus
try').show()
```

```
+----------------+----------+
|        industry|max_rating|
+----------------+----------+
|      Engineering|       760|
|       Technology|       750|
|Business Services|       780|
|    Manufacturing|       720|
+----------------+----------+
```

In [24]:
```
q1_df
```

Out[24]:
```
DataFrame[industry: string, max_rating: bigint]
```

In [26]:
```
q2_df = spark.sql('select parent_name, name, min(rating) as min_rating from prsd_json group
 by 1,2').show()
```

```
+--------------------+--------------------+----------+
|         parent_name|                name|min_rating|
+--------------------+--------------------+----------+
|KYOCERA UNIMERCO ...|KYOCERA UNIMERCO ...|       660|
|KYOCERA Document ...|KYOCERA Document ...|       730|
|KYOCERA UNIMERCO ...|KYOCERA UNIMERCO A/S|       620|
|                null|          KYOCERA Group|       490|
|KYOCERA Document ...|KYOCERA Document ...|       780|
|KYOCERA Document ...|KYOCERA Document ...|       770|
|KYOCERA Document ...|KYOCERA Document ...|       740|
|KYOCERA SENCO Ind...|KYOCERA SENCO Net...|       520|
|KYOCERA SENCO Net...|KYOCERA SENCO Deu...|       570|
|KYOCERA Document ...|KYOCERA Document ...|       730|
|KYOCERA Document ...|KYOCERA Document ...|       740|
|KYOCERA Document ...|KYOCERA Document ...|       590|
|KYOCERA Document ...|KYOCERA Document ...|       740|
|       KYOCERA Group|KYOCERA Fineceram...|       620|
|KYOCERA Document ...|KYOCERA Document ...|       650|
|KYOCERA Document ...|KYOCERA Document ...|       750|
|KYOCERA Document ...|KYOCERA Document ...|       550|
|       KYOCERA Group|KYOCERA Document ...|       510|
|KYOCERA Document ...|KYOCERA Document ...|       670|
|KYOCERA Document ...|KYOCERA Document ...|       740|
+--------------------+--------------------+----------+
only showing top 20 rows
```

In [28]:
```
q3_df = spark.sql('select distinct name, children_name as children, children_children_name a
s grand_children, children_children_children_name as great_grand_children from aa').show()
```

```
+------------+--------------------+--------------------+--------------------+
|        name|            children|      grand_children|great_grand_children|
+------------+--------------------+--------------------+--------------------+
|KYOCERA Group|KYOCERA Fineceram...|KYOCERA UNIMERCO ...|KYOCERA UNIMERCO A/S|
|KYOCERA Group|KYOCERA Fineceram...|KYOCERA UNIMERCO ...|KYOCERA UNIMERCO ...|
|KYOCERA Group|KYOCERA Document ...|KYOCERA Document ...|KYOCERA Document ...|
|KYOCERA Group|KYOCERA Document ...|KYOCERA Document ...|KYOCERA Document ...|
|KYOCERA Group|KYOCERA Document ...|KYOCERA Document ...|KYOCERA Document ...|
|KYOCERA Group|KYOCERA SENCO Ind...|KYOCERA SENCO Net...|    KYOCERA SENCO UK|
|KYOCERA Group|KYOCERA Document ...|KYOCERA Document ...|KYOCERA Document ...|
|KYOCERA Group|KYOCERA Document ...|KYOCERA Document ...|KYOCERA Document ...|
|KYOCERA Group|KYOCERA Document ...|KYOCERA Document ...|KYOCERA Document ...|
|KYOCERA Group|KYOCERA Fineceram...|KYOCERA UNIMERCO ...|KYOCERA UNIMERCO ...|
|KYOCERA Group|KYOCERA Document ...|KYOCERA Document ...|KYOCERA Document ...|
|KYOCERA Group|KYOCERA Document ...|KYOCERA Document ...|KYOCERA Document ...|
|KYOCERA Group|KYOCERA Document ...|KYOCERA Document ...|KYOCERA Document ...|
|KYOCERA Group|KYOCERA Document ...|KYOCERA Document ...|KYOCERA Document ...|
|KYOCERA Group|KYOCERA Document ...|KYOCERA Document ...|KYOCERA Document ...|
|KYOCERA Group|KYOCERA Fineceram...|KYOCERA UNIMERCO ...|KYOCERA UNIMERCO ...|
|KYOCERA Group|KYOCERA Fineceram...|KYOCERA UNIMERCO ...|KYOCERA UNIMERCO ...|
|KYOCERA Group|KYOCERA SENCO Ind...|KYOCERA SENCO Net...|KYOCERA SENCO Deu...|
|KYOCERA Group|KYOCERA Document ...|KYOCERA Document ...|KYOCERA Document ...|
|KYOCERA Group|KYOCERA Document ...|KYOCERA Document ...|KYOCERA Document ...|
+------------+--------------------+--------------------+--------------------+
only showing top 20 rows
```

In [ ]:
```
```