Which of the Spark API to use

Which of the Spark API to use

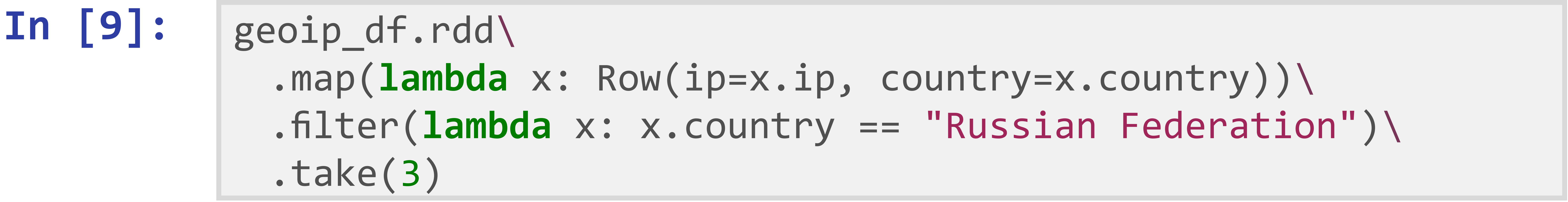
```
geoip_df
```

```
DataFrame[ip: string, code: string, country: string]
```

```
geoip_df.rdd.take(3)
```

```
[Row(ip=u'194.120.126.123', code=u'NL', country=u'Netherlands'),
Row(ip=u'94.126.119.173', code=u'FR', country=u'France'),
Row(ip=u'193.46.74.166', code=u'RU', country=u'Russian Federation')]
```

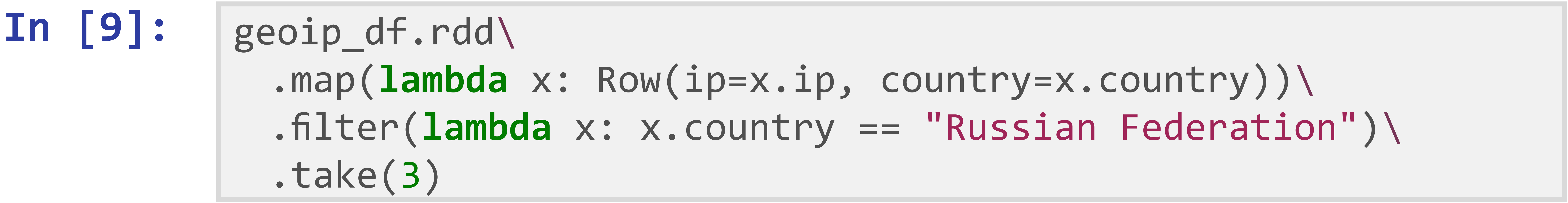
RDD API



In [9]:

```
geoip_df.rdd\
  .map(lambda x: Row(ip=x.ip, country=x.country))\
  .filter(lambda x: x.country == "Russian Federation")\
  .take(3)
```

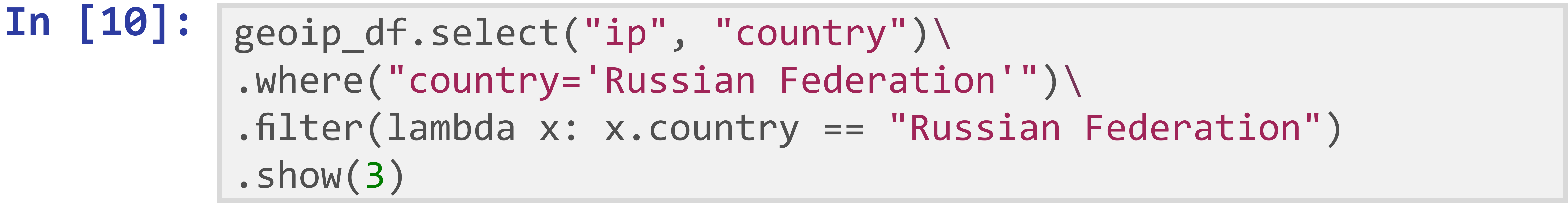
RDD API



In [9]:

```
geoip_df.rdd\
  .map(lambda x: Row(ip=x.ip, country=x.country))\
  .filter(lambda x: x.country == "Russian Federation")\
  .take(3)
```

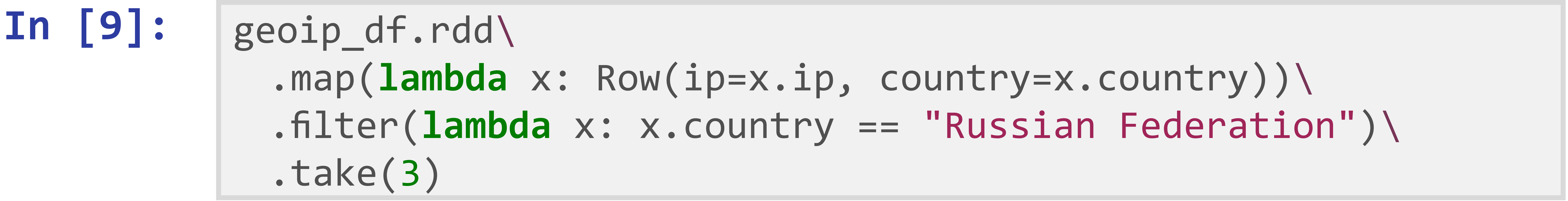
DataFrame API



In [10]:

```
geoip_df.select("ip", "country")\
.where("country='Russian Federation'")\
.filter(lambda x: x.country == "Russian Federation")
.show(3)
```

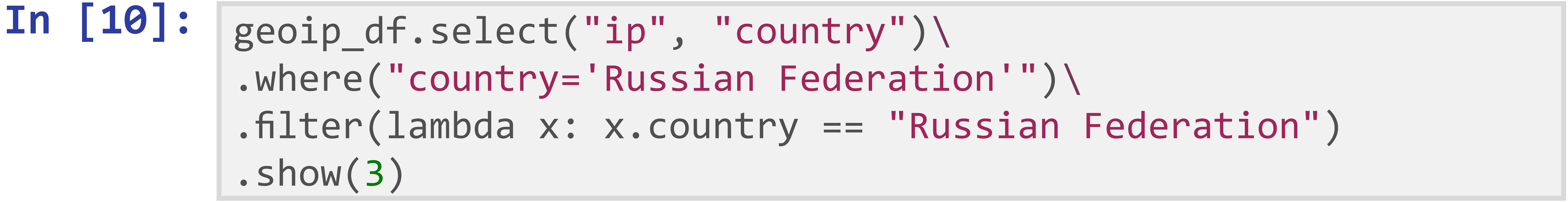
RDD API



In [9]:

```
geoip_df.rdd\
  .map(lambda x: Row(ip=x.ip, country=x.country))\
  .filter(lambda x: x.country == "Russian Federation")\
  .take(3)
```

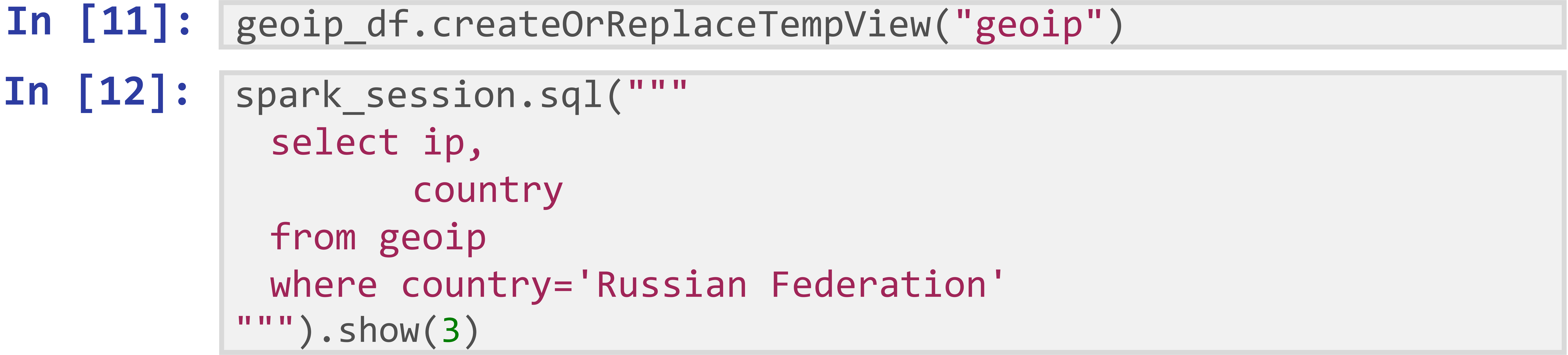
DataFrame API



In [10]:

```
geoip_df.select("ip", "country")\
.where("country='Russian Federation'")\
.filter(lambda x: x.country == "Russian Federation")
.show(3)
```

SQL



In [11]:

```
geoip_df.createOrReplaceTempView("geoip")
```

In [12]:

```
spark_session.sql("""
  select ip,
         country
  from geoip
  where country='Russian Federation'
""").show(3)
```

RDD API

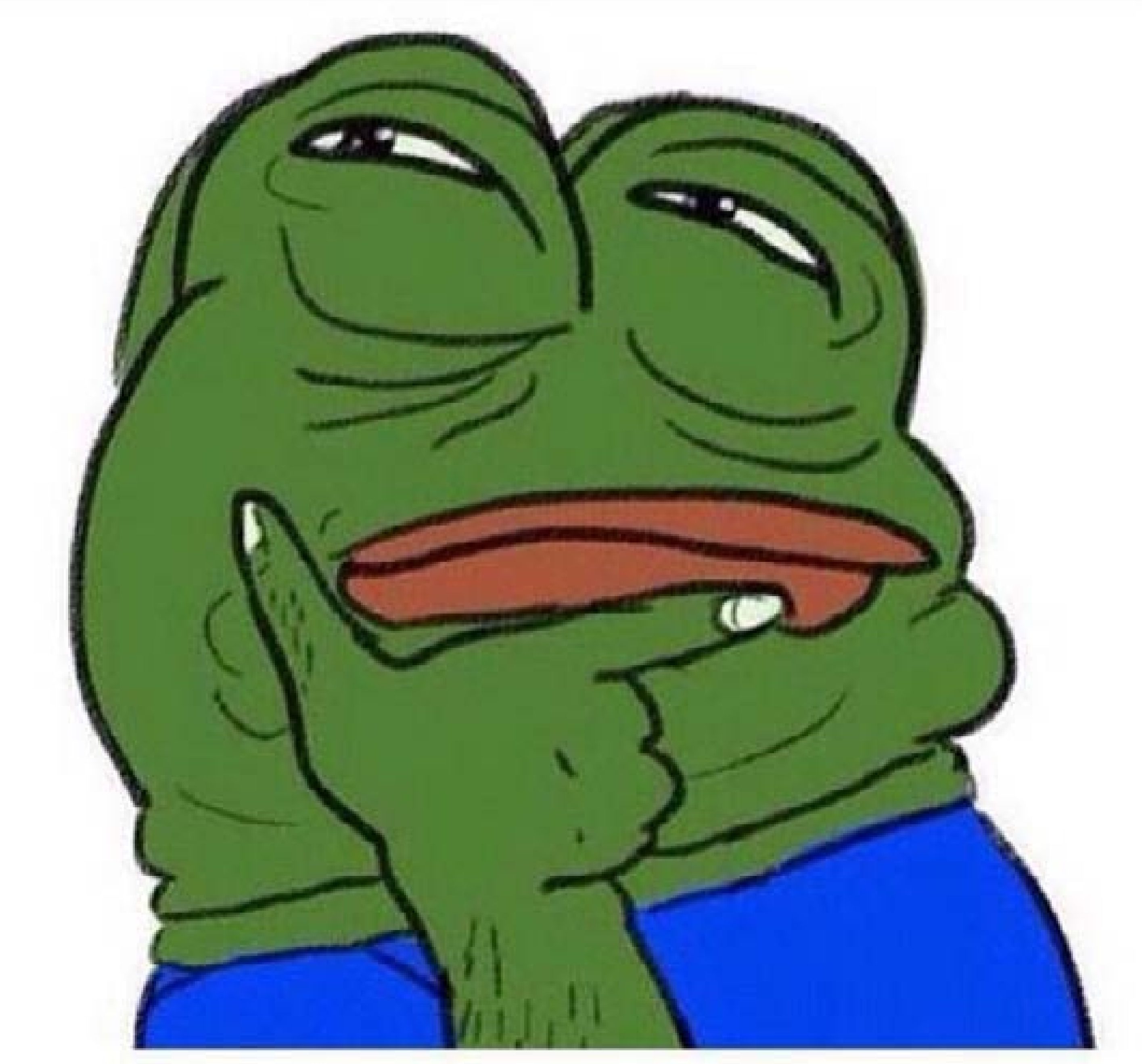
In [9]:

```
geoip_df.rdd\
  .map(lambda x: Row(ip=x.ip, country=x.country))\
  .filter(lambda x: x.country == "Russian Federation")\
  .take(3
```

DataFrame API

In [10]:

```
geoip_df.
.where("c                                  on'")\
.filter(la                                 ssian Federation")
.show(3)
```



SQL

In [11]:

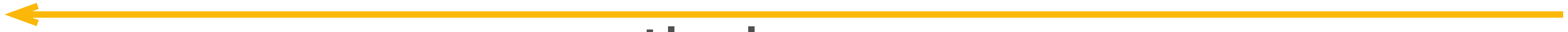
```
geoip_df.                                  "geoip")
```

In [12]:

```
spark_ses
  select ip,
         country
  from geoip
  where country='Russian Federation'
""").show(3)
```

RDD vs DataFrame & SQL

RDD vs DataFrame & SQL



optimizer

RDD vs DataFrame & SQL

←

optimizer

✗python

RDD vs DataFrame & SQL

optimizer

✗python

✓scala

RDD vs DataFrame & SQL

optimizer

RDD	DataFrame & SQL
✗python	✓scala
—	✓optimized code generation

RDD vs DataFrame & SQL

←

optimizer

✗python

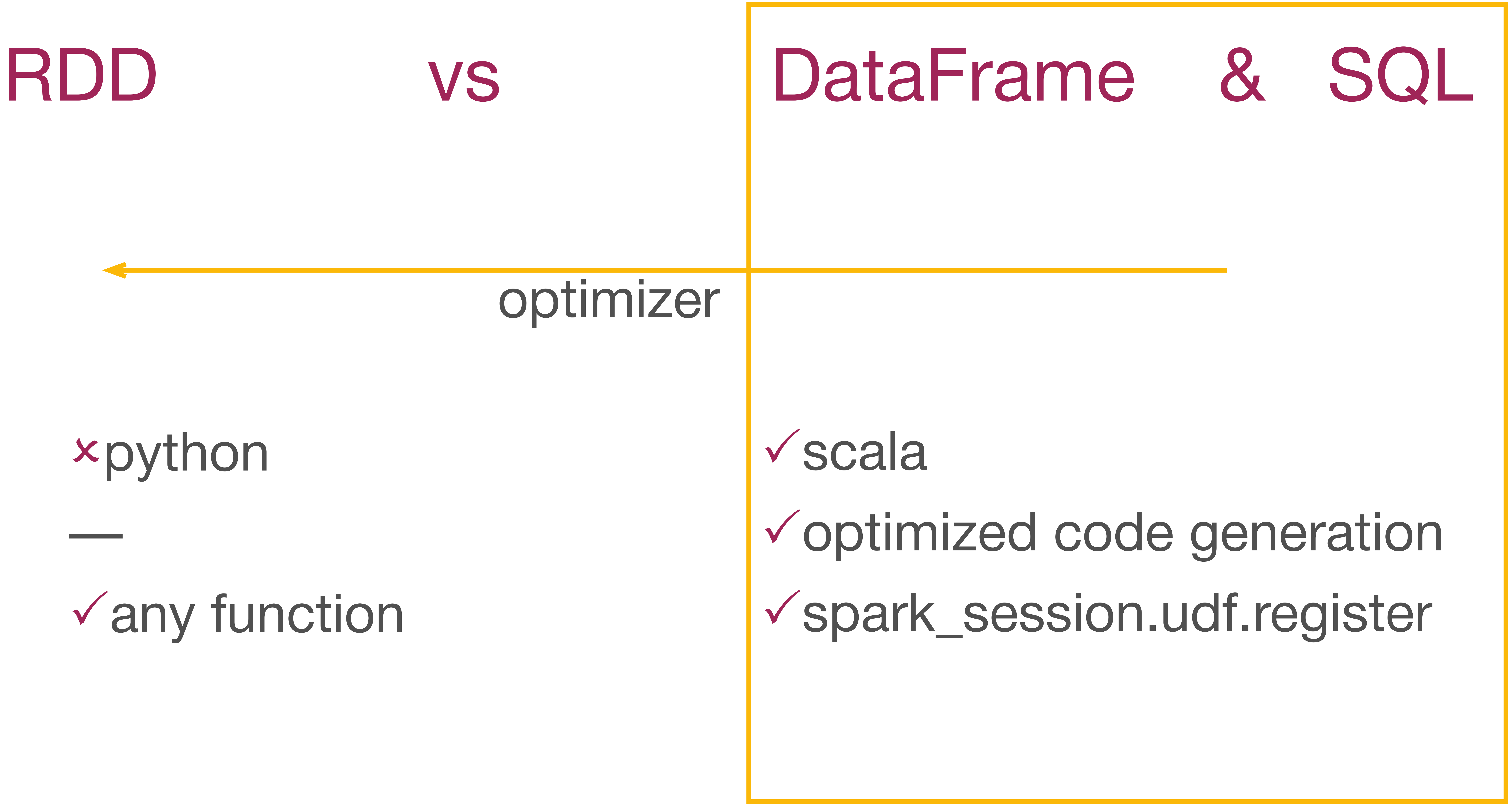
—

✓any function

✓scala

✓optimized code generation

✓spark_session.udf.register



RDD vs DataFrame & SQL

optimizer

✗python

—

✓any function

✓scala

✓optimized code generation

✓spark_session.udf.register

By DataFrame API vs By SQL command

```
geoip_df\
 .selec("ip","country")\
 .where("country='Russia'")\
 .show(3)
```

```
spark_session.sql("""
  selec ip,
         country
   from geoip
   where country='Russia'
""").show(3)
```

By DataFrame API vs By SQL command

```
geoip_df\
 .selec("ip","country")\
 .where("country='Russia'")\
 .show(3)
```

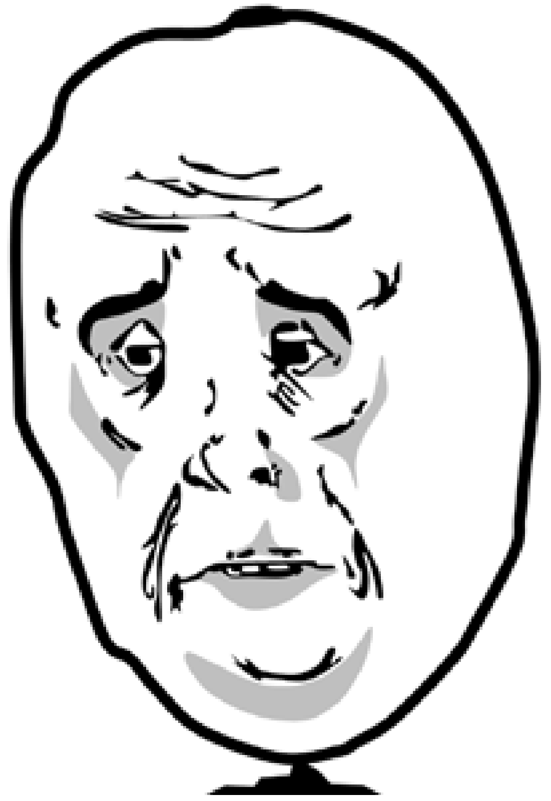
```
spark_session.sql("""
  selec ip,
         country
   from geoip
   where country='Russia'
""").show(3)
```

By DataFrame API vs By SQL command

```
geoip_df\
  .selec("ip","country")\
  .where("country='Russia'")\
  .show(3)
```

```
spark_session.sql("""
   selec ip,
           country
    from geoip
    where country='Russia'
""").show(3)
```

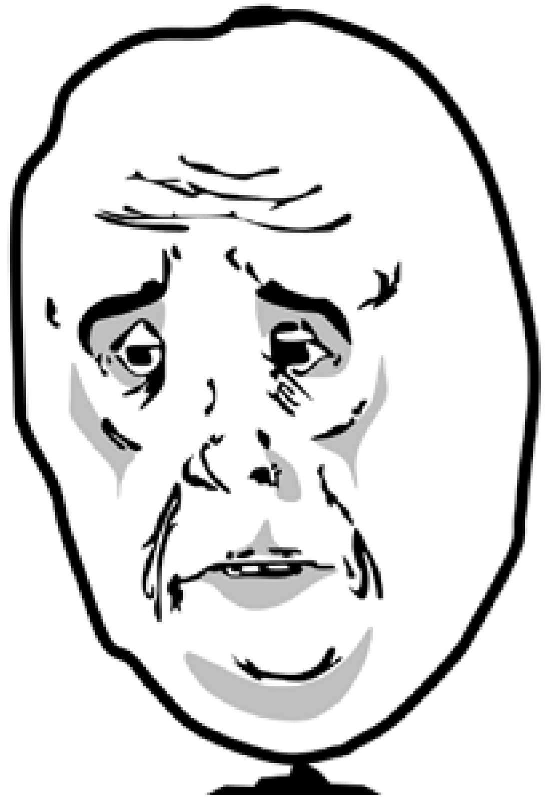
Error will be found at code compilation



By DataFrame API vs By SQL command

```
geoip_df\
 .selec("ip","country")\
 .where("country='Russia'")\
 .show(3)
```

Error will be found at code compilation



```
spark_session.sql("""
  selec ip,
         country
   from geoip
   where country='Russia'
""").show(3)
```

Error will be found at query call