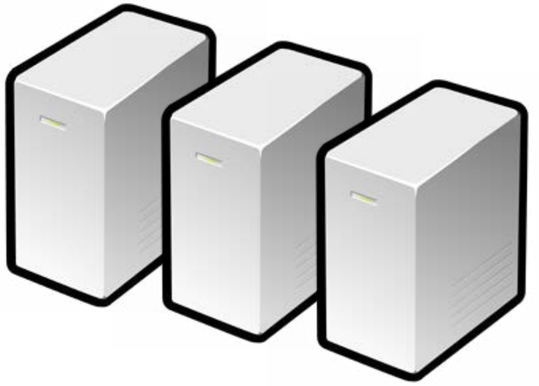
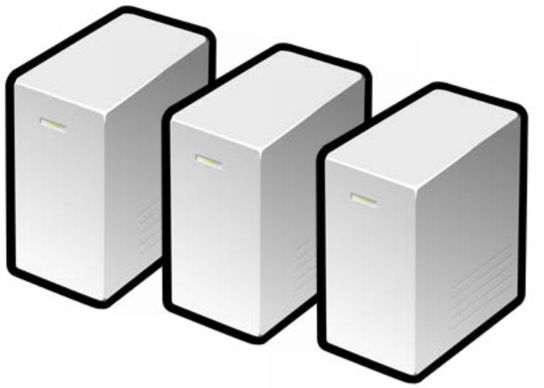
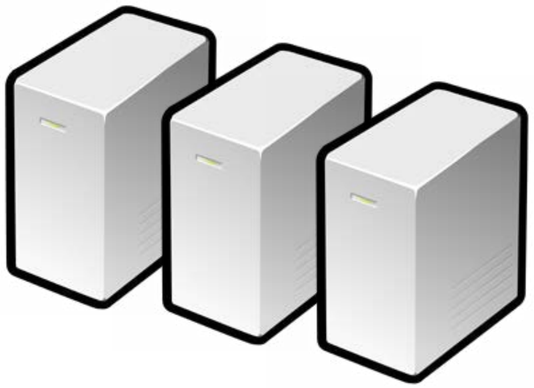
JOIN

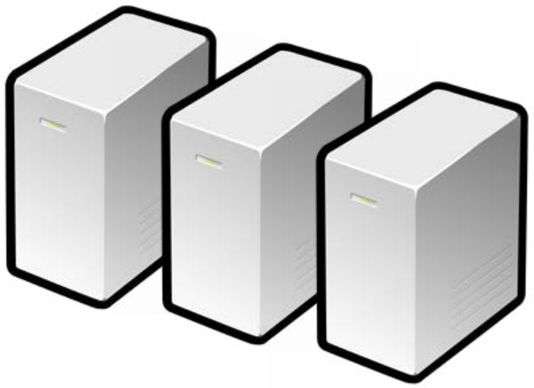




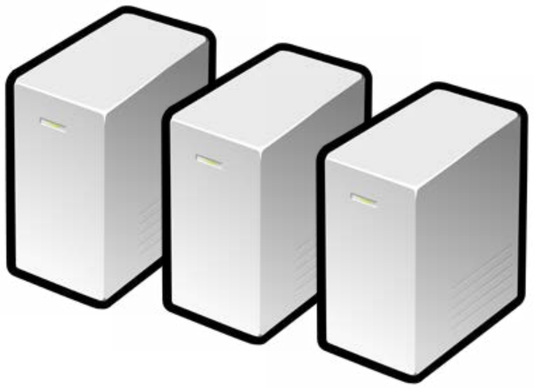


IP

IP

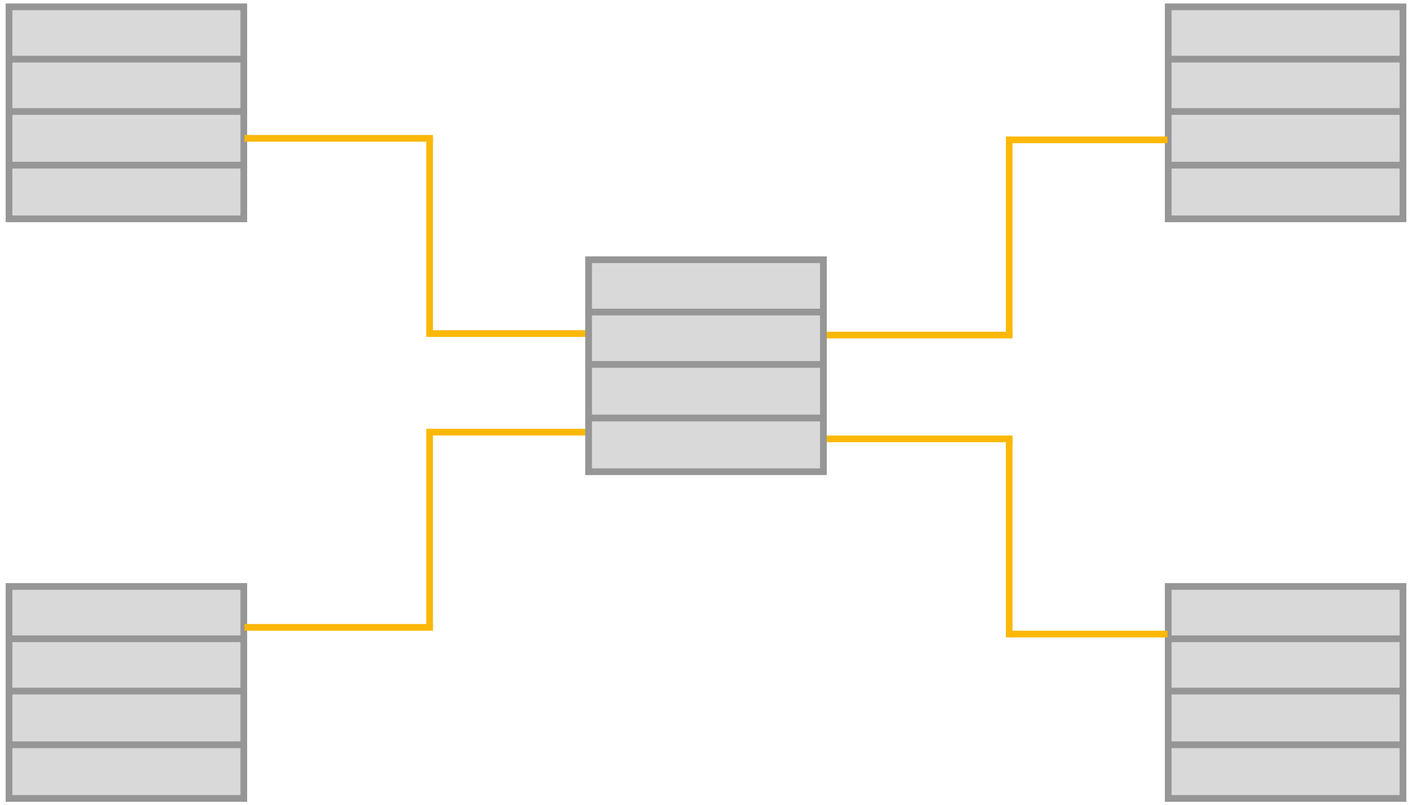


IP

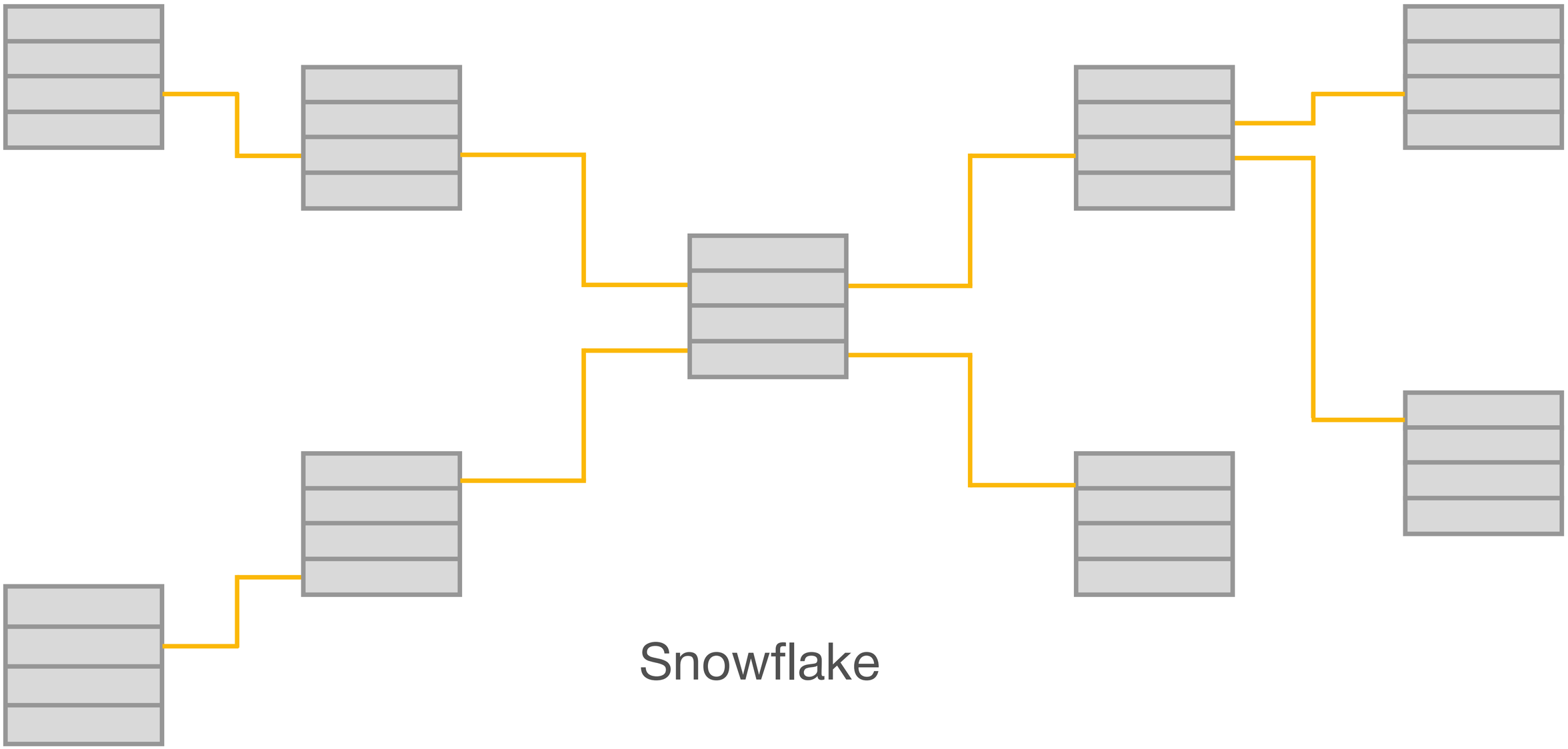


User IP

Join



Star



Snowflake


DSL
USB
Ethernet
4
3
2
1
LAN WAN
Power
12VDC
On
Off

Content

- How to join DataFrames
- Different types of join
- Why you need cross join

	Country	Sessions	% Sessions
1.	Russia	373	43.17%
2.	Germany	211	24.42%
3.	Ukraine	160	18.52%
4.	United States	26	3.01%
5.	(not set)	23	2.66%
6.	Belarus	17	1.97%

```
access_log = spark_session.read.table("web.access_log")
```

```
access_log.limit(3).toPandas()
```

	http_code	ip	response_length	time	url	user_agent
0	200	109.106.133.8	21546	12/Dec /2015:01:31:46 +0400	/id53821	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4)...
1	200	46.31.82.254	8777	12/Dec /2015:01:31:47 +0400	/id33929	Mozilla/5.0 (Windows NT 5.1; U; de; rv:1.9.1.6...
2	200	193.124.254.46	8731	12/Dec /2015:01:31:48 +0400	/id35754	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT...

```
geoip = spark_session.read.table("web.geoip")
```

```
geoip.limit(3).toPandas()
```

	http_code	ip	code	country
0	200	194.120.126.123	NL	Netherlands
1	200	94.126.119.173	FR	France
2	200	193.46.74.166	RU	Russian Federation

Join

```
spark_session.sql("""
select *
from web.access_log l
join web.geoip g
on l.ip = g.ip
""").limit(3).toPandas()
```

	http_code	ip	response_length	time	url	user_agent	code	country
0	200	109.106.133.8	21546	12/Dec /2015:01:31:46 +0400	/id53821	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4)...	RU	Russian Federation
1	200	46.31.82.254	8777	12/Dec /2015:01:31:47 +0400	/id33929	Mozilla/5.0 (Windows NT 5.1; U; de; rv:1.9.1.6...	GB	United Kingdom
2	200	193.124.254.46	8731	12/Dec /2015:01:31:48 +0400	/id35754	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT...	RU	Russian Federation

```
spark_session.sql("""
select *
from web.access_log l
join web.geoip g
on l.ip = g.ip
""").limit(3).toPandas()
```

	http_code	ip	response_length	time	url	user_agent	code	country
0	200	109.106.133.8	21546	12/Dec /2015:01:31:46 +0400	/id53821	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4)...	RU	Russian Federation
1	200	46.31.82.254	8777	12/Dec /2015:01:31:47 +0400	/id33929	Mozilla/5.0 (Windows NT 5.1; U; de; rv:1.9.1.6...	GB	United Kingdom
2	200	193.124.254.46	8731	12/Dec /2015:01:31:48 +0400	/id35754	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT...	RU	Russian Federation

```
spark_session.sql("""
select *
from web.access_log l
join web.geoip g
on l.ip = g.ip
""").limit(3).toPandas()
```

	http_code	ip	response_length	time	url	user_agent	code	country
0	200	109.106.133.8	21546	12/Dec /2015:01:31:46 +0400	/id53821	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4)...	RU	Russian Federation
1	200	46.31.82.254	8777	12/Dec /2015:01:31:47 +0400	/id33929	Mozilla/5.0 (Windows NT 5.1; U; de; rv:1.9.1.6...	GB	United Kingdom
2	200	193.124.254.46	8731	12/Dec /2015:01:31:48 +0400	/id35754	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT...	RU	Russian Federation

```
spark_session.sql("""
select *
from web.access_log l
join web.geoip g
on l.ip = g.ip
""").limit(3).toPandas()
```

	http_code	ip	response_length	time	url	user_agent	code	country
0	200	109.106.133.8	21546	12/Dec /2015:01:31:46 +0400	/id53821	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4)...	RU	Russian Federation
1	200	46.31.82.254	8777	12/Dec /2015:01:31:47 +0400	/id33929	Mozilla/5.0 (Windows NT 5.1; U; de; rv:1.9.1.6...	GB	United Kingdom
2	200	193.124.254.46	8731	12/Dec /2015:01:31:48 +0400	/id35754	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT...	RU	Russian Federation

accesslog

```
spark_session.sql("""
select *
from web.access_log l
join web.geoip g
on l.ip = g.ip
""").limit(3).toPandas()
```

	http_code	ip	response_length	time	url	user_agent	code	country
0	200	109.106.133.8	21546	12/Dec /2015:01:31:46 +0400	/id53821	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4)...	RU	Russian Federation
1	200	46.31.82.254	8777	12/Dec /2015:01:31:47 +0400	/id33929	Mozilla/5.0 (Windows NT 5.1; U; de; rv:1.9.1.6...	GB	United Kingdom
2	200	193.124.254.46	8731	12/Dec /2015:01:31:48 +0400	/id35754	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT...	RU	Russian Federation

accesslog

geo

dataframe.join

```
access_log.join(geoip_df, on="ip")\
          .limit(3).toPandas()
```

```
access_log.join(geoip_df, on="ip")\
          .limit(3).toPandas()
```

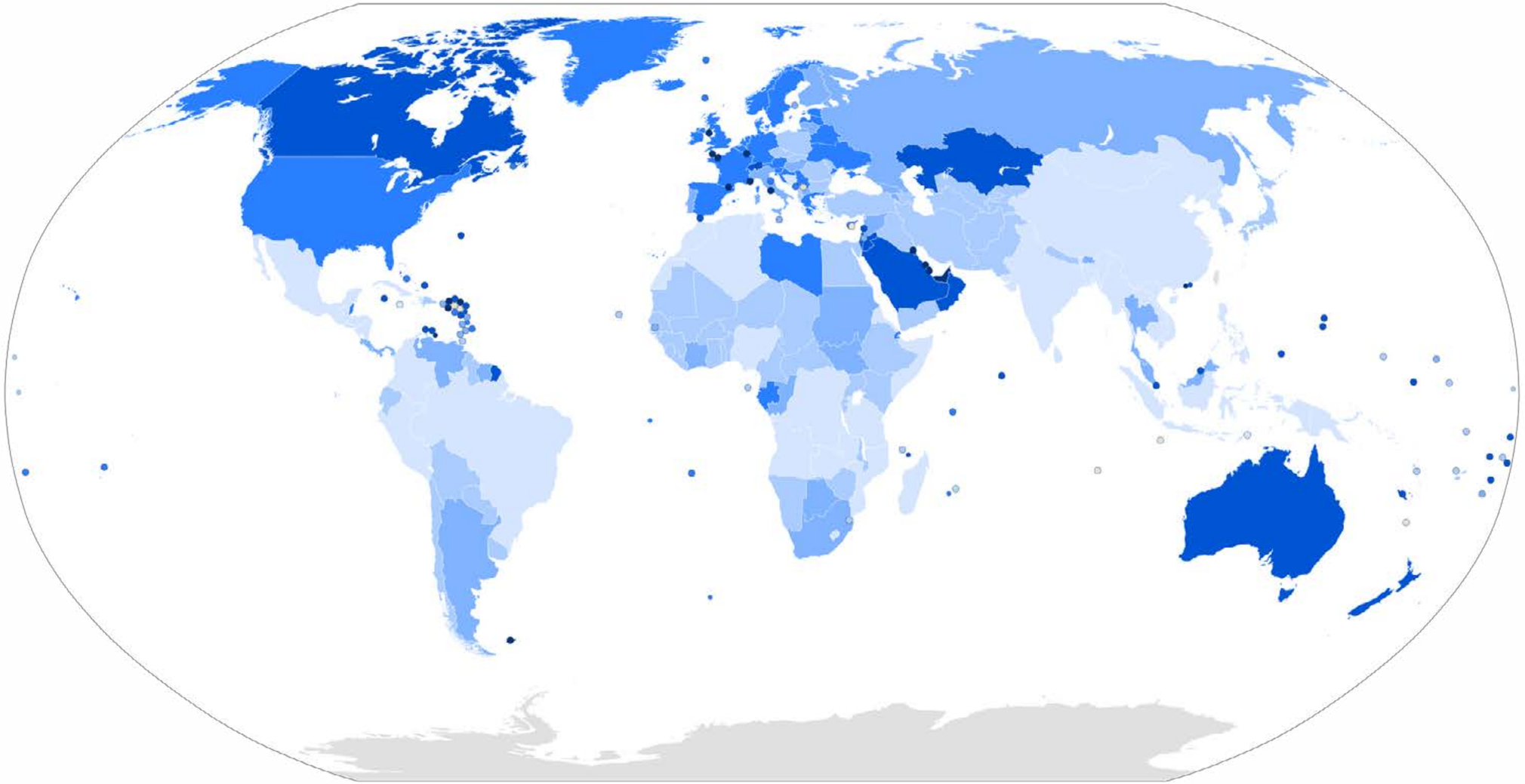
```
access_log.join(geoip_df, on="ip")\
          .limit(3).toPandas()
```

```
access_log.join(geoip_df, on="ip")\
          .limit(3).toPandas()
```

	http_code	ip	response_length	time	url	user_agent	code	country
0	200	109.106.133.8	21546	12/Dec /2015:01:31:46 +0400	/id53821	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4)...	RU	Russian Federation
1	200	46.31.82.254	8777	12/Dec /2015:01:31:47 +0400	/id33929	Mozilla/5.0 (Windows NT 5.1; U; de; rv:1.9.1.6...	GB	United Kingdom
2	200	193.124.254.46	8731	12/Dec /2015:01:31:48 +0400	/id35754	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT...	RU	Russian Federation

```
access_log.join(geoip_df,
                on = access_log.ip == geoip_df.ip)\
          .limit(3).toPandas()
```

	http_code	ip	response_length	time	url	user_agent	code	country
0	200	109.106.133.8	21546	12/Dec /2015:01:31:46 +0400	/id53821	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4)...	RU	Russian Federation
1	200	46.31.82.254	8777	12/Dec /2015:01:31:47 +0400	/id33929	Mozilla/5.0 (Windows NT 5.1; U; de; rv:1.9.1.6...	GB	United Kingdom
2	200	193.124.254.46	8731	12/Dec /2015:01:31:48 +0400	/id35754	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT...	RU	Russian Federation



```
access_log.join(geoip_df, on="ip")\
          .limit(3).toPandas()
```

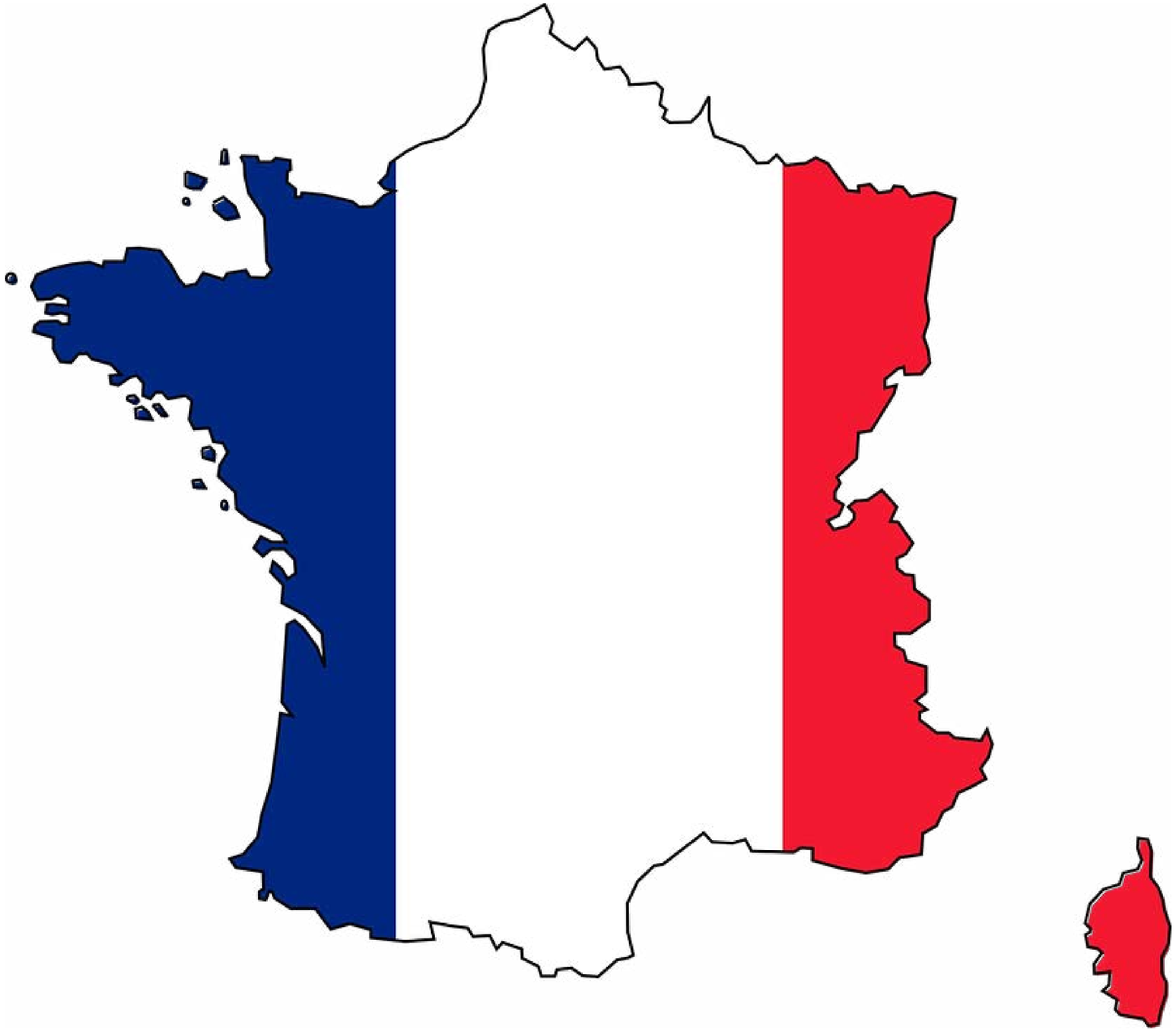
	http_code	ip	response_length	time	url	user_agent	code	country
0	200	109.106.133.8	21546	12/Dec /2015:01:31:46 +0400	/id53821	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4)...	RU	Russian Federation
1	200	46.31.82.254	8777	12/Dec /2015:01:31:47 +0400	/id33929	Mozilla/5.0 (Windows NT 5.1; U; de; rv:1.9.1.6...	GB	United Kingdom
2	200	193.124.254.46	8731	12/Dec /2015:01:31:48 +0400	/id35754	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT...	RU	Russian Federation

```
access_log.join(geoip_df, on = "ip",)\
          .groupby("country")\
          .agg(f.countDistinct("ip").alias("cnt"))\
          .limit(3).toPandas()
```

	country	cnt
0	Sweden	247
1	Singapore	5
2	Turkey	2

```
access_log.join(geoip_df, on = "ip",)\
          .groupby("country")\
          .agg(f.countDistinct("ip").alias("cnt"))\
          .orderBy(f.col("cnt").desc())\
          .limit(3).toPandas()
```

	country	cnt
0	Russian Federation	4556
1	France	1474
2	Germany	1287



```
geoip.count()
```

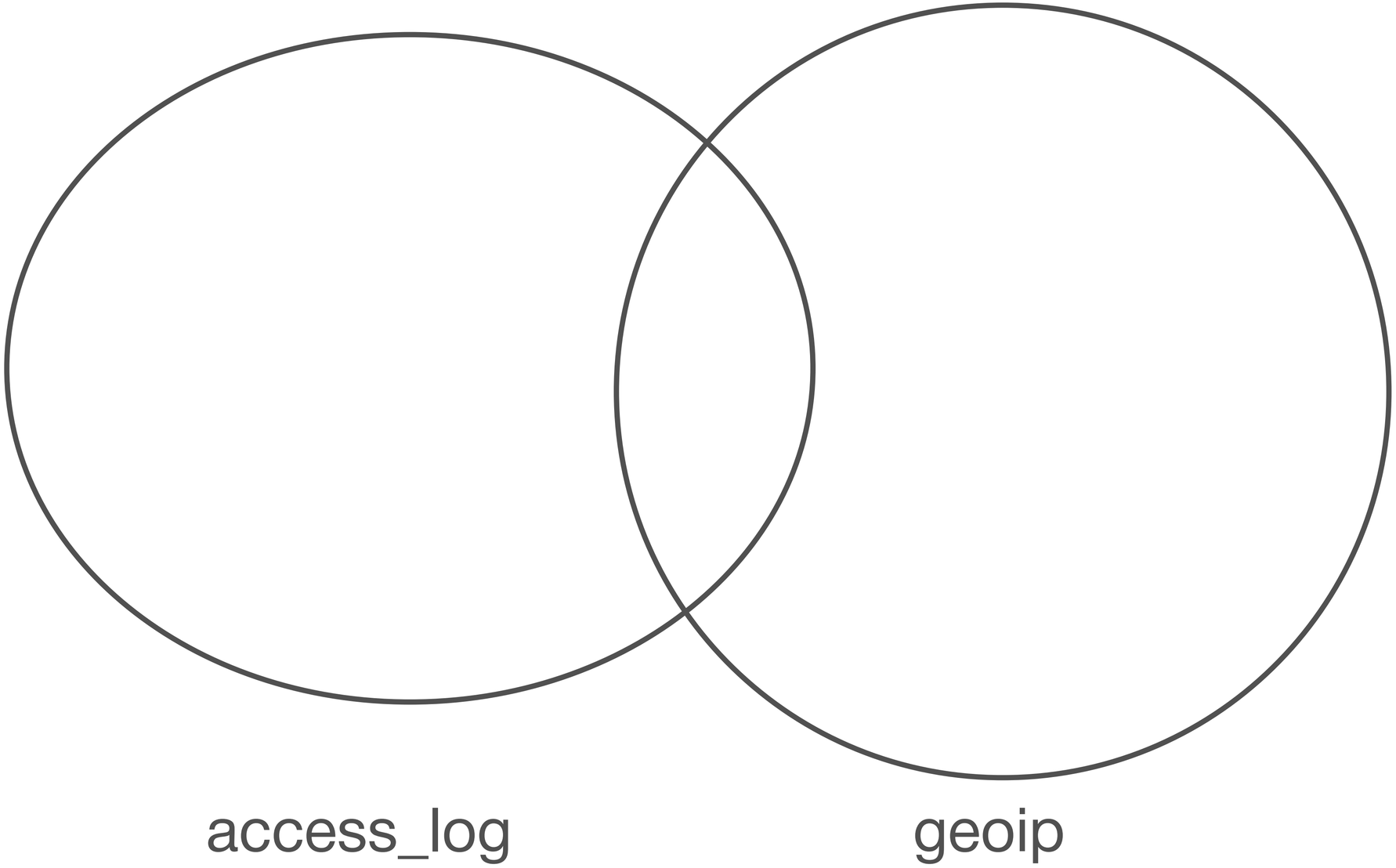
9910

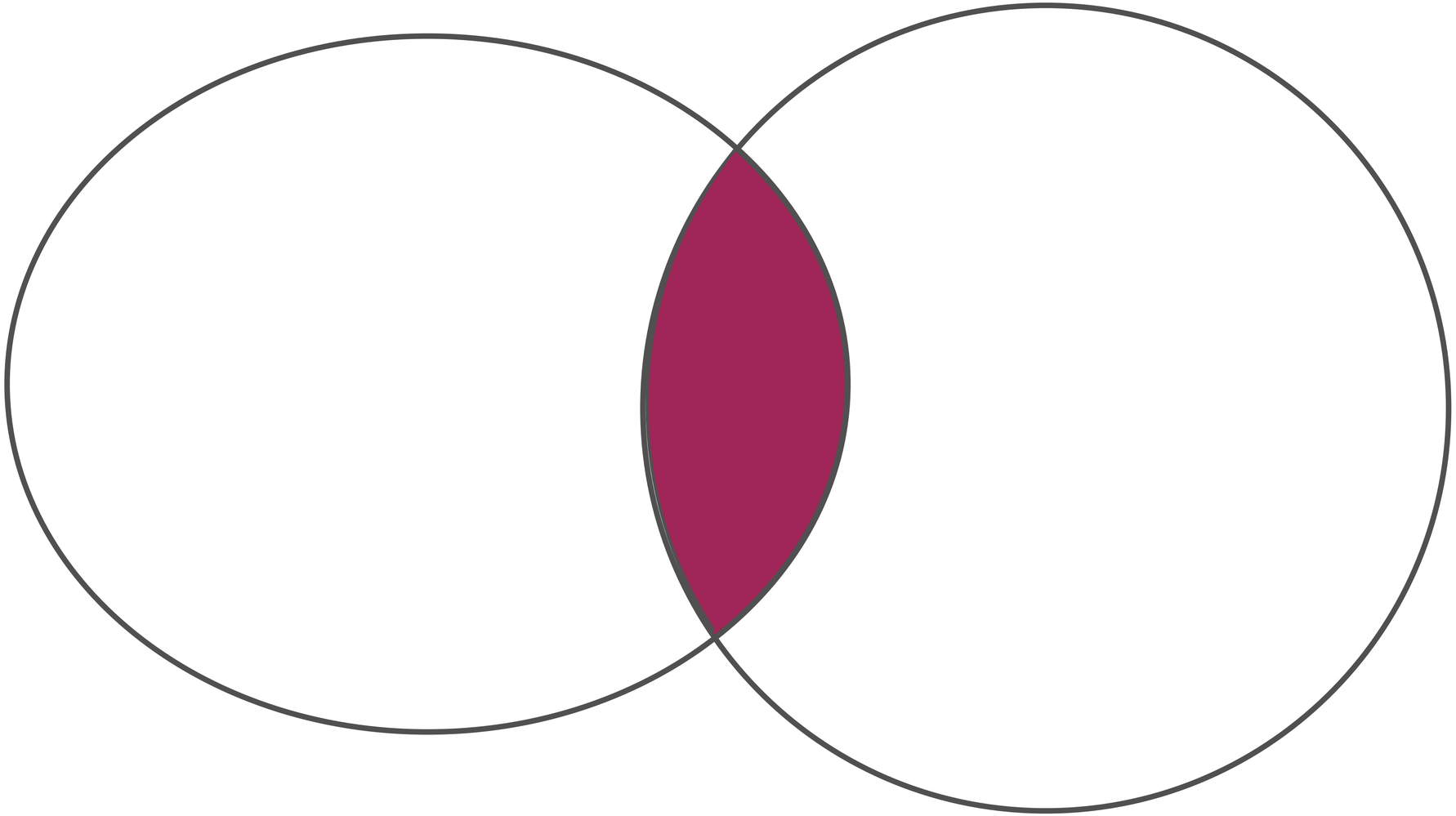
```
access_log.count()
```

89206

```
access_log.join(geoip_df, on = "ip",)\
           .count()
```

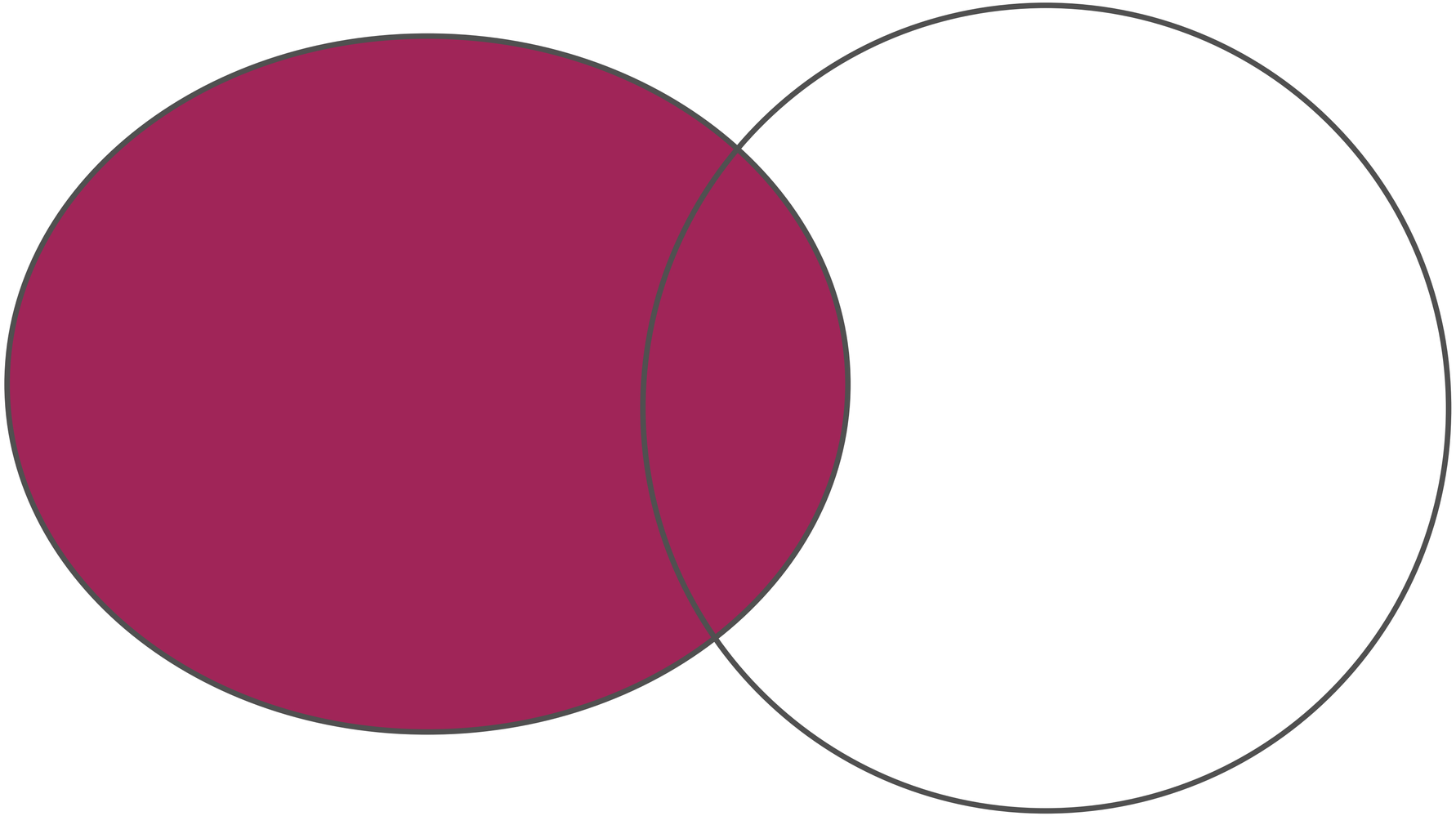
88774


access_log
geoip



```
access_log.join(geoip_df,
                on = "ip",
                how = "inner")\
          .limit(3).toPandas()
```

	http_code	ip	response_length	time	url	user_agent	code	country
0	200	109.106.133.8	21546	12/Dec /2015:01:31:46 +0400	/id53821	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4)...	RU	Russian Federation
1	200	46.31.82.254	8777	12/Dec /2015:01:31:47 +0400	/id33929	Mozilla/5.0 (Windows NT 5.1; U; de; rv:1.9.1.6...	GB	United Kingdom
2	200	193.124.254.46	8731	12/Dec /2015:01:31:48 +0400	/id35754	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT...	RU	Russian Federation



```
access_log.join(geoip_df,
                on = "ip",
                how = "left")\
           .count()
```

89206

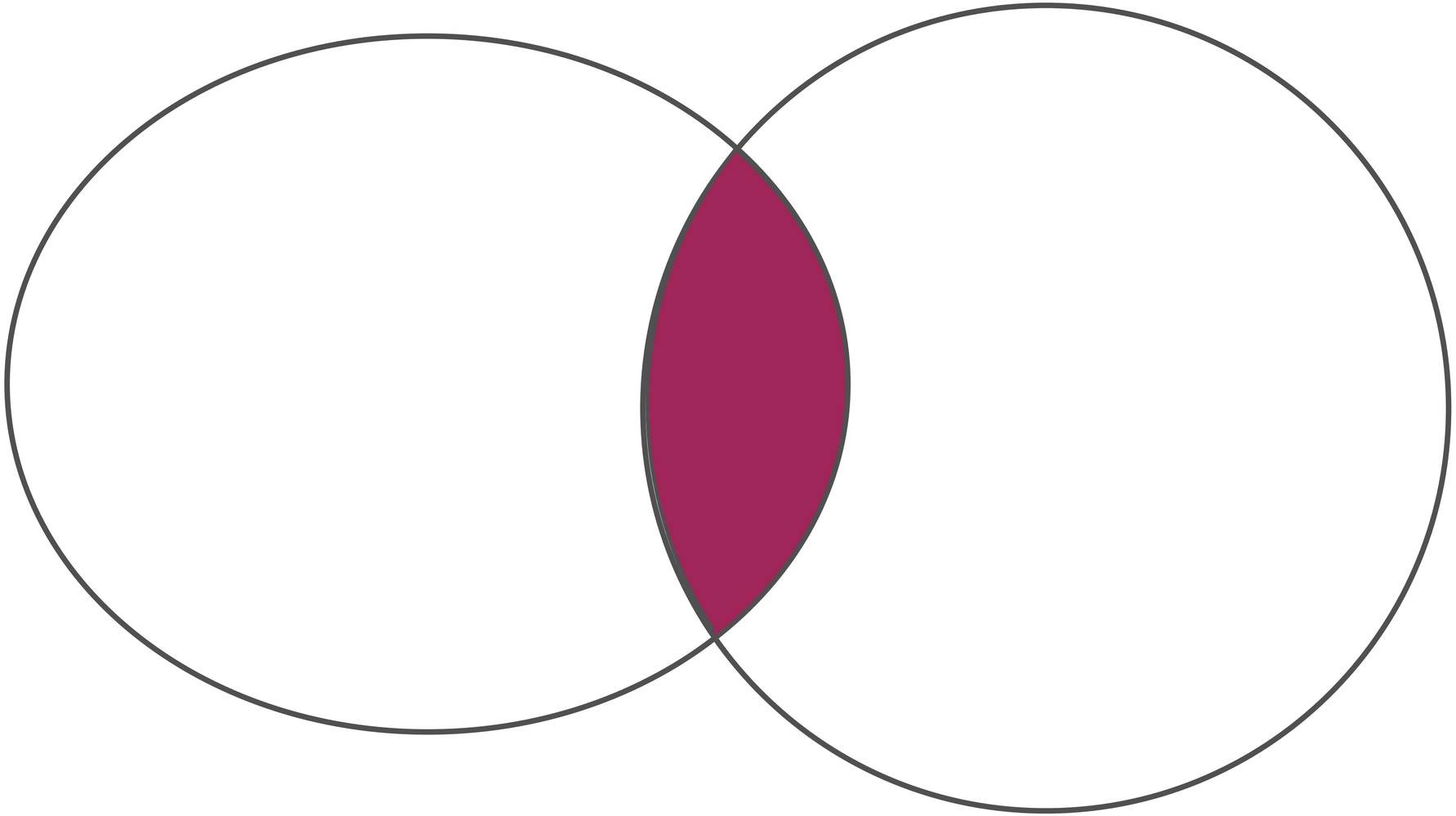
Semi join

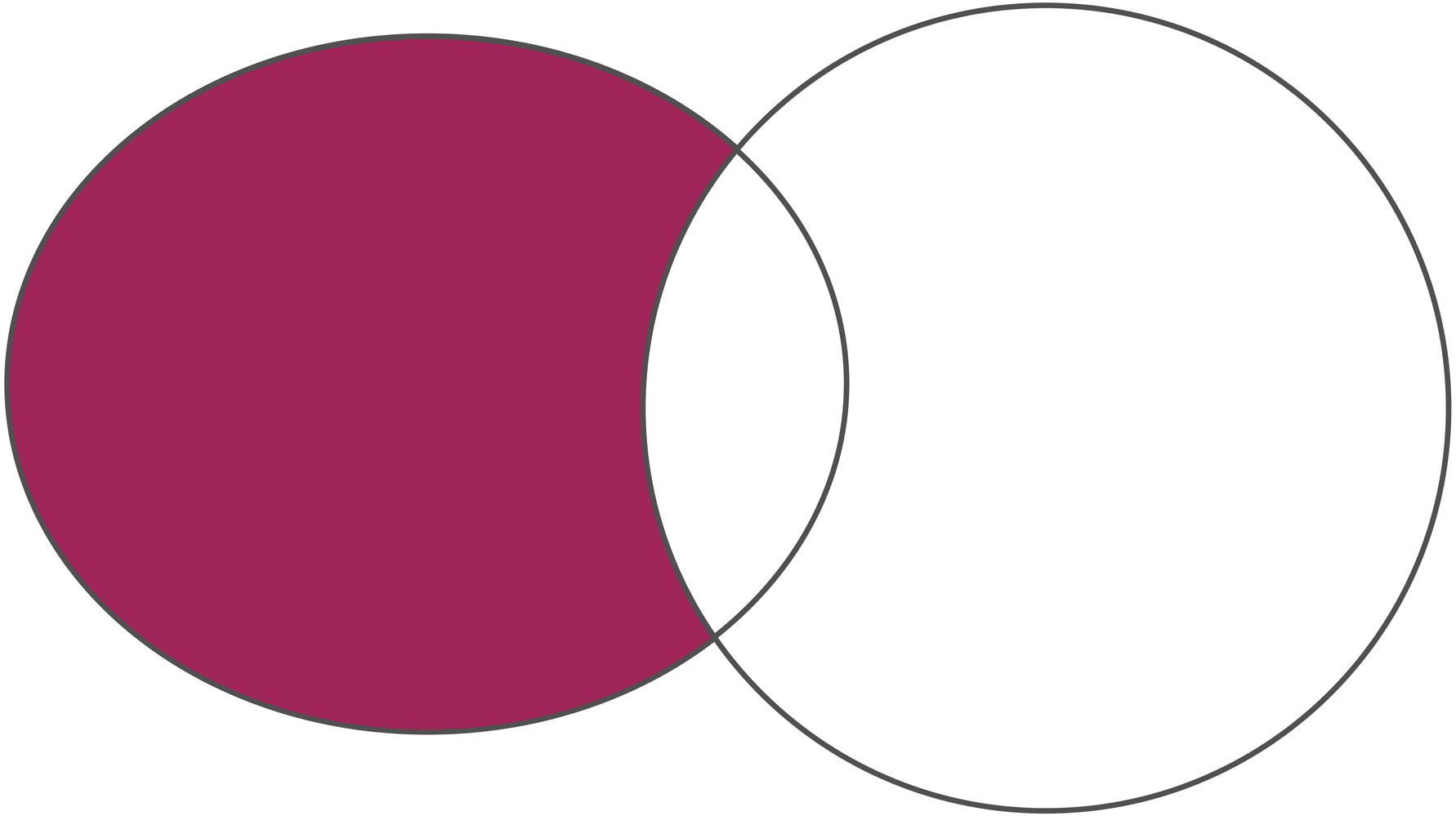
```
access_log.join(geoip_df,
                on = "ip",
                how = "left_semi")\
          .limit(3).toPandas()
```

	ip	http_code	response_length	time	url	user_agent
0	109.106.133.8	200	21546	12/Dec /2015:01:31:46 +0400	/id53821	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4)...
1	46.31.82.254	200	8777	12/Dec /2015:01:31:47 +0400	/id33929	Mozilla/5.0 (Windows NT 5.1; U; de; rv:1.9.1.6...
2	193.124.254.46	200	8731	12/Dec /2015:01:31:48 +0400	/id35754	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT...

```
access_log.join(geoip_df,
            on = "ip",
            how = "left_semi")\
        .limit(3).toPandas()
```

88774





```
access_log.join(geoip_df,
                on = "ip",
                how = "left_anti")\
          .limit(3).toPandas()
```

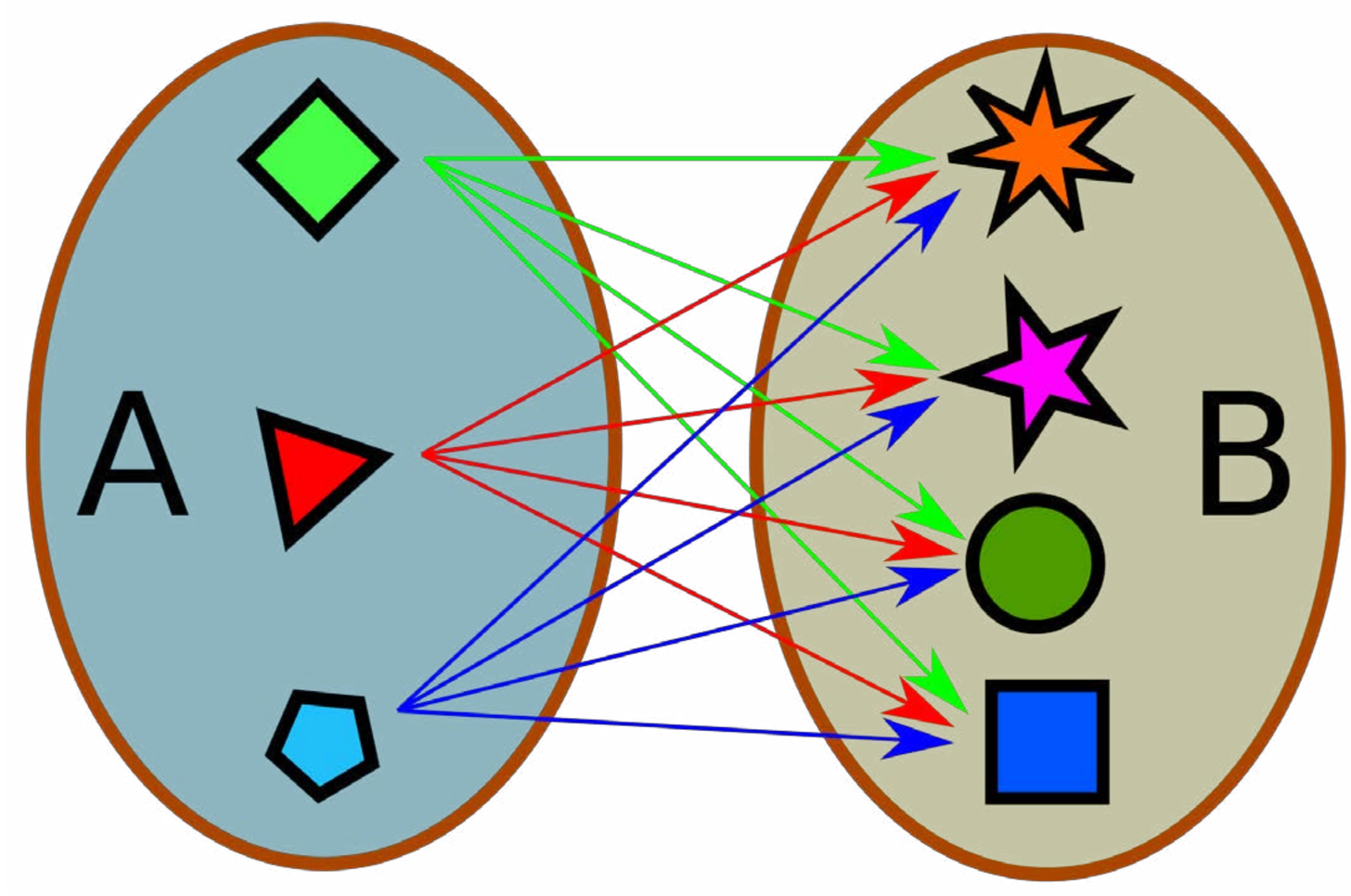
	ip	http_code	response_length	time	url	user_agent
0	197.189.56.86	404	0	12/Dec /2015:01:32:17 +0400	/admin.php	Mozilla/6.66
1	91.212.123.110	200	13193	12/Dec /2015:01:33:20 +0400	/	Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKi...
2	91.212.123.110	404	0	12/Dec /2015:01:33:20 +0400	/favicon.ico	Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKi...

```
access_log.join(geoip_df,
            on = "ip",
            how = "left_anti")\
           .count()
```

432

Cross join

Cross join



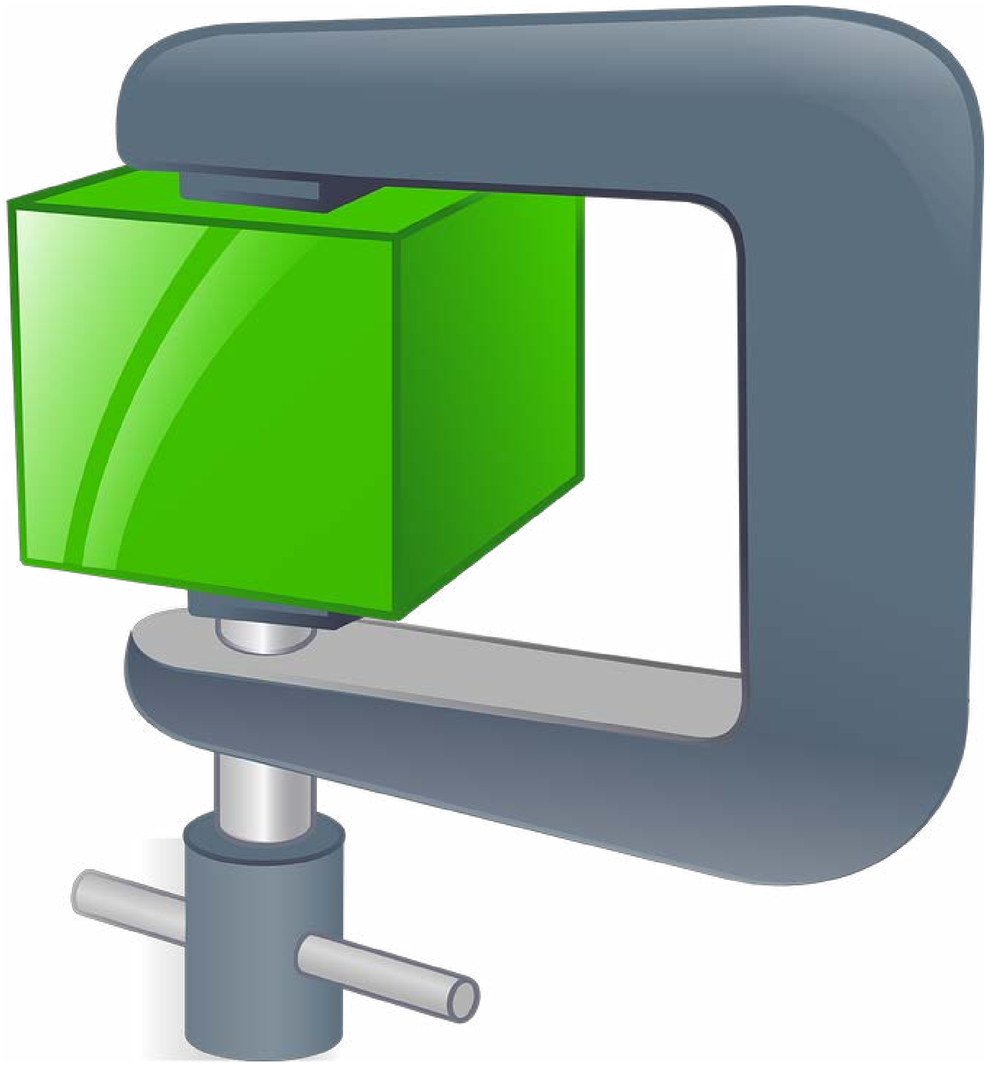
```
access_log.crossJoin(geoip_df)\
           .count()
```

884031460

1x1=1	2x1=2	3x1=3	4x1=4	5x1=5
1x2=2	2x2=4	3x2=6	4x2=8	5x2=10
1x3=3	2x3=6	3x3=9	4x3=12	5x3=15
1x4=4	2x4=8	3x4=12	4x4=16	5x4=20
1x5=5	2x5=10	3x5=15	4x5=20	5x5=25
1x6=6	2x6=12	3x6=18	4x6=24	5x6=30
1x7=7	2x7=14	3x7=21	4x7=28	5x7=35
1x8=8	2x8=16	3x8=24	4x8=32	5x8=40
1x9=9	2x9=18	3x9=27	4x9=36	5x9=45
1x10=10	2x10=20	3x10=30	4x10=40	5x10=50

6x1=6	7x1=7	8x1=8	9x1=9	10x1=10
6x2=12	7x2=14	8x2=16	9x2=18	10x2=20
6x3=18	7x3=21	8x3=24	9x3=27	10x3=30
6x4=24	7x4=28	8x4=32	9x4=36	10x4=40
6x5=30	7x5=35	8x5=40	9x5=45	10x5=50
6x6=36	7x6=42	8x6=48	9x6=54	10x6=60
6x7=42	7x7=49	8x7=56	9x7=63	10x7=70
6x8=48	7x8=56	8x8=64	9x8=72	10x8=80
6x9=54	7x9=63	8x9=72	9x9=81	10x9=90
6x10=60	7x10=70	8x10=80	9x10=90	10x10=100

Log squashing



```
squashed_log = access_log.select("*",
                 (f.rand()*10+1).astype("int").alias("events_count"))
aquashed_log.limit(5).toPandas()
```

http_code		ip	response_length	time	url	user_agent	events_count
0	200	109.106.133.8	21546	12/Dec /2015:01:31:46 +0400	/id53821	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4)...	8
1	200	46.31.82.254	8777	12/Dec /2015:01:31:47 +0400	/id33929	Mozilla/5.0 (Windows NT 5.1; U; de; rv:1.9.1.6...	2
2	200	193.124.254.46	8731	12/Dec /2015:01:31:48 +0400	/id35754	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT...	2
3	200	185.103.220.164	22776	12/Dec /2015:01:31:48 +0400	/id78231	Mozilla/5.0 (Linux;Android 4.4.2; nb-no;SAMS...	3
4	200	185.103.220.164	18335	12/Dec /2015:01:31:48 +0400	/id39395	Mozilla/5.0 (Linux; Android 4.4.2; nb-no; SAMS...	3

```python
import pandas as pd
ids_pd = pd.DataFrame({"id":range(10)})
ids_pd
```

	id
0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9

```
ids = spark_session.createDataFrame(ids_pd)
```

```
squashed_log.crossJoin(ids)\
              .orderBy("ip", "url", "time", "id")\
              .limit(5).toPandas()
```

	http_code	ip	response_length	time	url	user_agent	events_count	id
0	200	100.43.79.212	30116	10/Dec/2015...	\	Mozilla/5.0 (Wind...	6	0
1	200	100.43.79.212	30116	10/Dec/2015...	\	Mozilla/5.0 (Wind...	6	1
2	200	100.43.79.212	30116	10/Dec/2015...	\	Mozilla/5.0 (Wind...	6	2
3	200	100.43.79.212	30116	10/Dec/2015...	\	Mozilla/5.0 (Wind...	6	3
4	200	100.43.79.212	30116	10/Dec/2015...	\	Mozilla/5.0 (Wind...	6	4

```
squashed_log.crossJoin(ids)\
            .count()
```

892060

```
squashed_log.crossJoin(ids)\
              .where(f.col("id") < f.col("events_count"))\
              .orderBy("ip", "url", "time", "id")\
              .limit(5).toPandas()
```

	http_code	ip	response_length	time	url	user_agent	events_count	id
0	200	100.43.79.212	30116	10/Dec/2015...	\	Mozilla/5.0 (Wind...	6	0
1	200	100.43.79.212	30116	10/Dec/2015...	\	Mozilla/5.0 (Wind...	6	1
2	200	100.43.79.212	30116	10/Dec/2015...	\	Mozilla/5.0 (Wind...	6	2
3	200	100.43.79.212	30116	10/Dec/2015...	\	Mozilla/5.0 (Wind...	6	3
4	200	100.43.79.212	30116	10/Dec/2015...	\	Mozilla/5.0 (Wind...	6	4
5	200	100.43.79.212	30116	10/Dec/2015...	\	Mozilla/5.0 (Wind...	6	5
6	404	100.43.79.212	0	10/Dec/2015...	\favicon.	Mozilla/5.0 (Wind...	3	0
7	404	100.43.79.212	0	10/Dec/2015...	\favicon.	Mozilla/5.0 (Wind...	3	1

```
squashed_log.crossJoin(ids)\
            .where(f.col("id") <= f.col ("events_count"))\
            .count()
```

571012

You have learned

- how to join on a DataFrame
- the differences of join types
- the trick of data processing using cross join