

Two-Dimensional Distributions

```
access_log.limit(5).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....
2	200	193.124.254.46	8731	12/Dec/2015 :01:31:48 +0400	/id35754	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT ...
3	200	185.103.220.164	22776	12/Dec/2015 :01:31:48 +0400	/id78231	Mozilla/5.0 (Linux; Android 4.4.2;...
4	200	185.103.220.164	18335	12/Dec/2015 :01:31:49 +0400	/id39395	Mozilla/5.0 (Linux; Android 4.4.2;...

```
access_log.limit(5).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....
2	200	193.124.254.46	8731	12/Dec/2015:01:31:48+0400	/id35754	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT ...
3	200	185.103.220.164	22776	12/Dec/2015:01:31:48+0400	/id78231	Mozilla/5.0 (Linux; Android 4.4.2;...
4	200	185.103.220.164	18335	12/Dec/2015:01:31:49+0400	/id39395	Mozilla/5.0 (Linux; Android 4.4.2;...

country

RF

GBr

Fr

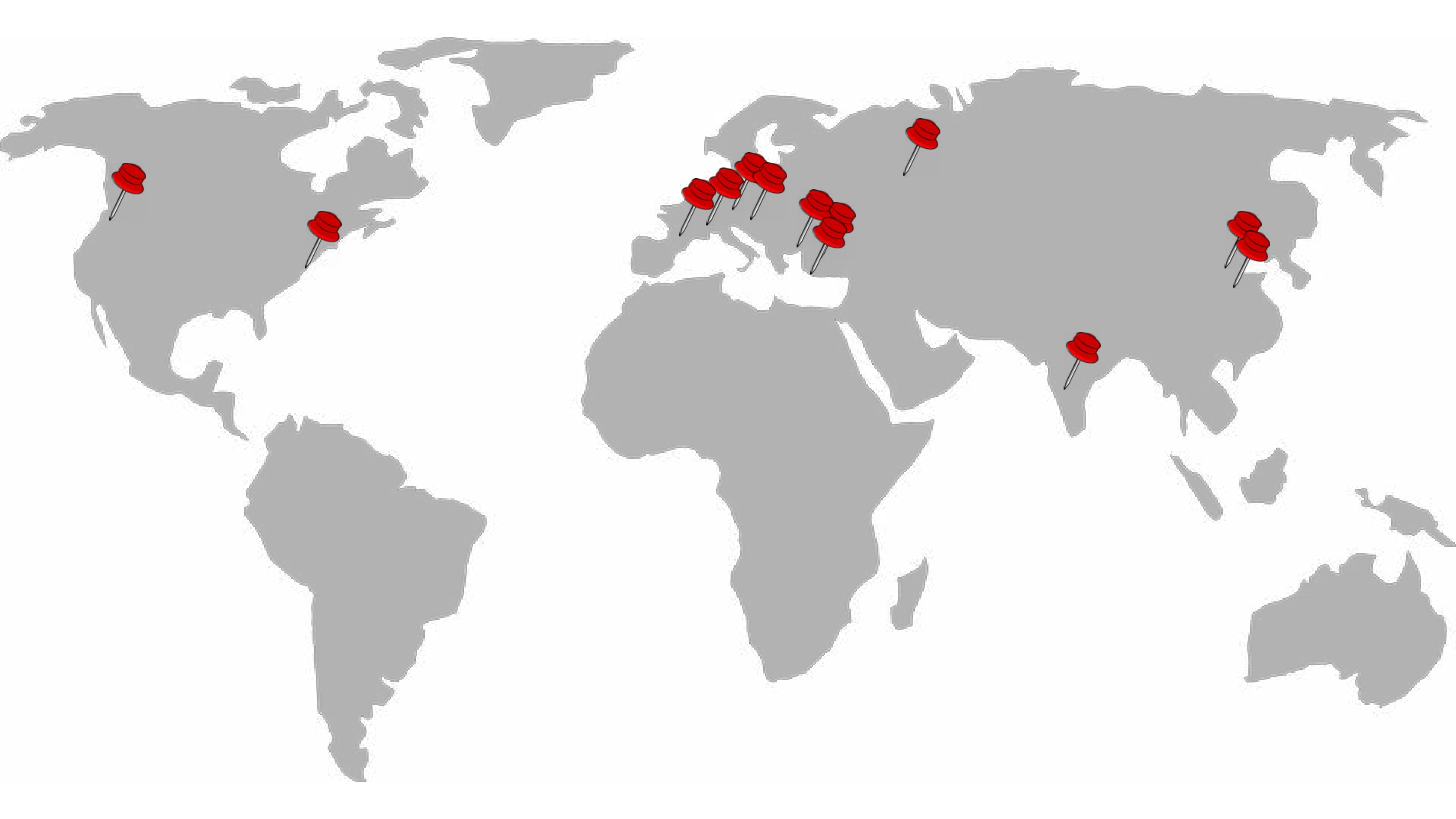
Ger

Ru

```
access_log.limit(5).toPandas()
```

	http_code	ip	response_length	time	url	user_agent	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)...	RF
1	200	46.31.82.254	8777	12/Dec/2015:01:31:47+0400	/id33929	Mozilla/5.0 (Windows NT 5.1....	GBr
2	200	193.124.254.46	8731	12/Dec/2015:01:31:48+0400	/id35754	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT ...	Fr
3	200	185.103.220.164	22776	12/Dec/2015:01:31:48+0400	/id78231	Mozilla/5.0 (Linux; Android 4.4.2;...	Ger
4	200	185.103.220.164	18335	12/Dec/2015:01:31:49+0400	/id39395	Mozilla/5.0 (Linux; Android 4.4.2;...	Ru

	ger	fr	...	rf	gbr
url 1	10%	0%	...	30%	5%
url 2	5%	20%	...	5%	10%
url 3	25%	50%	...	0%	10%
...



```
access_log.limit(5).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....
2	200	193.124.254.46	8731	12/Dec/2015:01:31:48+0400	/id35754	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT ...
3	200	185.103.220.164	22776	12/Dec/2015:01:31:48+0400	/id78231	Mozilla/5.0 (Linux; Android 4.4.2;...
4	200	185.103.220.164	18335	12/Dec/2015:01:31:49+0400	/id39395	Mozilla/5.0 (Linux; Android 4.4.2;...

```
access_log.limit(5).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....
2	200	193.124.254.46	8731	12/Dec/2015:01:31:48+0400	/id35754	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT ...
3	200	185.103.220.164	22776	12/Dec/2015:01:31:48+0400	/id78231	Mozilla/5.0 (Linux; Android 4.4.2;...
4	200	185.103.220.164	18335	12/Dec/2015:01:31:49+0400	/id39395	Mozilla/5.0 (Linux; Android 4.4.2;...


```
access_log.limit(5).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....
2	200	193.124.254.46	8731	12/Dec/2015:01:31:48+0400	/id35754	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT ...
3	200	185.103.220.164	22776	12/Dec/2015:01:31:48+0400	/id78231	Mozilla/5.0 (Linux; Android 4.4.2;...
4	200	185.103.220.164	18335	12/Dec/2015:01:31:49+0400	/id39395	Mozilla/5.0 (Linux; Android 4.4.2;...

	url1	url2	...	url999	url1000
user1	0	0	...	1	0
user 2	0	2	...	0	1
user 3	1	0	...	0	0
...




```
access_log.where("url =' /id73370' ")\  
    .groupBy("ip")\  
    .count()\  
    .limit(5).toPandas()
```

	ip	count
0	51.255.50.116	1
1	188.191.80.160	1
2	148.56.182.160	1
3	194.69.134.239	1
4	83.149.11.33	1

```
access_log.where("url =' /id73370' ")\  
  .groupBy("ip")\  
  .count()\  
  .limit(5).toPandas()
```

	ip	count
0	51.255.50.116	1
1	188.191.80.160	1
2	148.56.182.160	1
3	194.69.134.239	1
4	83.149.11.33	1

join

```
access_log.where("url =' /id73324' ")\  
  .groupBy("ip")\  
  .count()\  
  .limit(5).toPandas()
```

	ip	count
0	185.79.58.36	1
1	176.117.18.92	1
2	23.50.136.126	1
3	5.57.220.125	1
4	31.148.245.54	1

```
access_log.where("url = '/id73370'")\
    .groupBy("ip")\
    .count()\
    .limit(5).toPandas()
```

```
access_log.where("url = '/id73324'")\
    .groupBy("ip")\
    .count()\
    .limit(5).toPandas()
```

	ip	count
0	51.255.50.116	1
1	188.191.80.160	1
2	148.56.182.160	1
3	194.69.134.239	1
4	83.149.11.33	1

join

x1000

	ip	count
0	185.79.58.36	1
1	176.117.18.92	1
2	23.50.136.126	1
3	5.57.220.125	1
4	31.148.245.54	1

In this video you will learn

- what pivot table is
- how to build tables 2 dimensional distributions by it




```
access_log.where\  
    .groupBy("url")\  
    .count()\  
    .limit(5).toPandas()
```

	url	count
0	/id37020	4
1	/id47695	2
2	/id77559	1
3	/id47124	2
4	/id74109	2

```
access_log.where\  
    .groupBy("url")\  
    .count()\  
    .orderBy(f.col("count").desc())\  
    .limit(5).toPandas()
```

	url	count
0	/	13590
1	/favicon.ico	13590
2	/id91370	7
3	/id67946	6
4	/id73324	6

```
access_log.where\  
    .groupBy("url")\  
    .count()\  
    .orderBy(f.col("count").desc())\  
    .limit(1000).toPandas()  
  
top_url_pd.head(5)
```

	url	count
0	/	13590
1	/favicon.ico	13590
2	/id91370	7
3	/id73324	6
4	/id67946	6

```
top_url_list = topUrl["url"].tolist()
```

```
top_url_list[:5]
```

```
[u'/', u'/favicon.ico', u'/id91370', u'/id73324', u'/id67946']
```

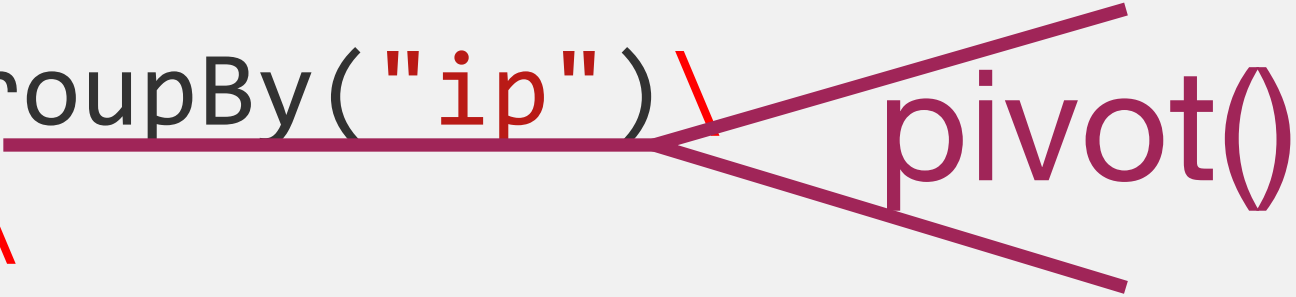


```
access_log.groupby("ip")\  
    .count()\  
    .limit(5).toPandas()
```

	ip	count
0	62.187.94.9	4
1	78.68.118.92	3
2	217.197.14.94	4
3	37.128.131.92	3
4	193.84.177.216	6

`pivot (column,[values])`

```
access_log.groupby("ip")\
    .count()\
    .limit(5).toPandas()
```



	ip	count
0	62.187.94.9	4
1	78.68.118.92	3
2	217.197.14.94	4
3	37.128.131.92	3
4	193.84.177.216	6


```
access_log.groupBy("ip")\  
  .pivot("url", top_url_list)\  
  .count()\  
  .limit(5).toPandas()
```

	ip	/	/favicon.ico	/id91370	...	/id36818	/id96232	/id41751	/id61427
0	23.54.65.93	1	1	None	...	None	None	None	None
1	158.69.69.225	2	2	None	...	None	None	None	None
2	81.163.111.217	2	2	None	...	None	None	None	None
3	178.208.51.84	3	3	None	...	None	None	None	None
4	109.60.192.67	1	1	None	...	None	None	None	None

5 rows × 1001 columns

```
access_log.groupBy("ip")\  
  .pivot("url", top_url_list)\  
  .count()\  
  .limit(5).toPandas()
```

	ip	/	/favicon.ico	/id91370	...	/id36818	/id96232	/id41751	/id61427
0	23.54.65.93	1	1	None	...	None	None	None	None
1	158.69.69.225	2	2	None	...	None	None	None	None
2	81.163.111.217	2	2	None	...	None	None	None	None
3	178.208.51.84	3	3	None	...	None	None	None	None
4	109.60.192.67	1	1	None	...	None	None	None	None

5 rows × 1001 columns

```
access_log.groupBy("ip")\
    .pivot("url", top_url_list)\
    .count()\
    .limit(5).toPandas()
```

	ip	/	/favicon.ico	/id91370	...	/id36818	/id96232	/id41751	/id61427
0	23.54.65.93	1	1	None	...	None	None	None	None
1	158.69.69.225	2	2	None	...	None	None	None	None
2	81.163.111.217	2	2	None	...	None	None	None	None
3	178.208.51.84	3	3	None	...	None	None	None	None
4	109.60.192.67	1	1	None	...	None	None	None	None

5 rows × 1001 columns

```
access_log.groupby("ip")\
    .pivot("url", top_url_list)\
    .count()\
    .fillna(0)\
    .limit(5).toPandas()
```

	ip	/	/favicon.ico	/id91370	...	/id36818	/id96232	/id41751	/id61427
0	23.54.65.93	1	1	0	...	0	0	0	0
1	158.69.69.225	2	2	0	...	0	0	0	0
2	81.163.111.217	2	2	0	...	0	0	0	0
3	178.208.51.84	3	3	0	...	0	0	0	0
4	109.60.192.67	1	1	0	...	0	0	0	0

5 rows × 1001 columns

Summary

- what pivot table is
- how to build tables 2 dimensional distributions by it

