Functions


f(x)

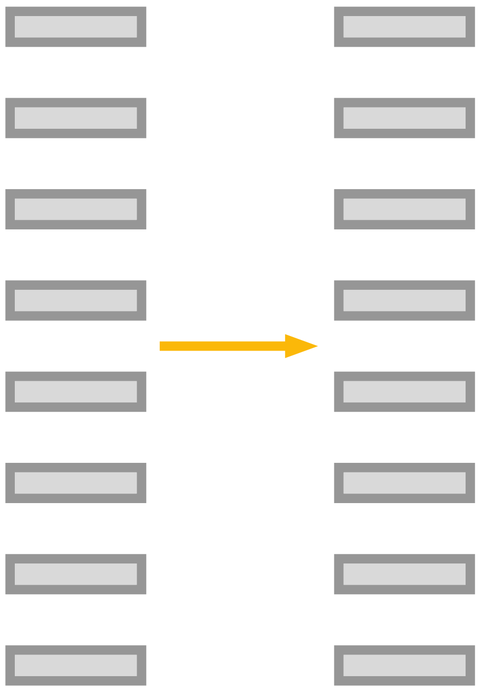
```
spark_session.sql("""
   select user_agent,
          length (user_agent)
   from web.access_log
   limit 3
   """).toPandas()
```

```
spark_session.sql("""
   select user_agent,
         length (user_agent)
   from web.access_log
   limit 3
   """).toPandas()
```

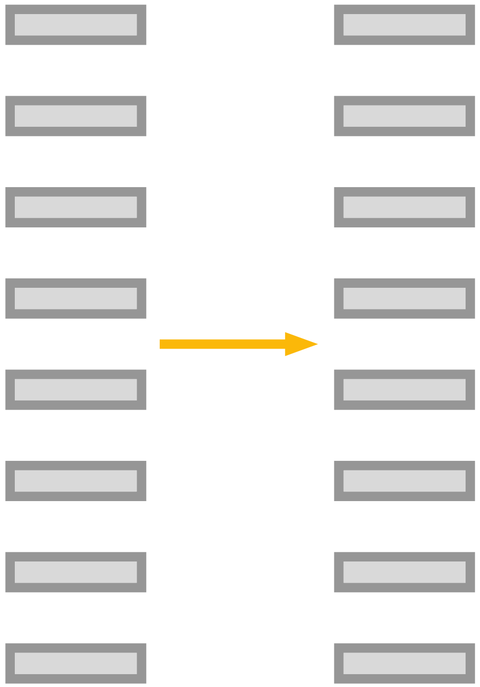
	user_agent	length (user_agent)
0	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4)...	120
1	Mozilla/5.0 (Windows NT 5.1; U; de; rv:1.9.1.6...	88
2	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT...	153

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...

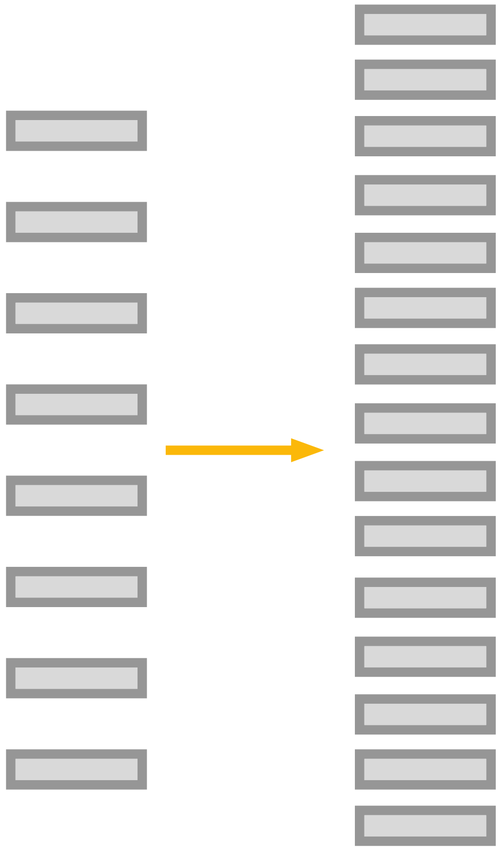
http_code		ip	response_length	time	url	user_agent
0	200	109.106.133.8	21546	12/Dec /2015:01:31:46 +0400	/id53821	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4)...
1	200	46.31.82.254	8777	12/Dec /2015:01:31:47 +0400	/id33929	Mozilla/5.0 (Windows NT 5.1; U; de; rv:1.9.1.6...
2	200	193.124.254.46	8731	12/Dec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



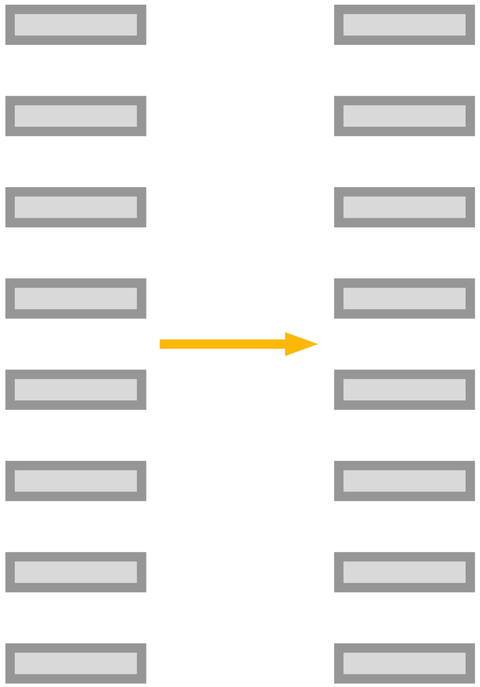
mapping



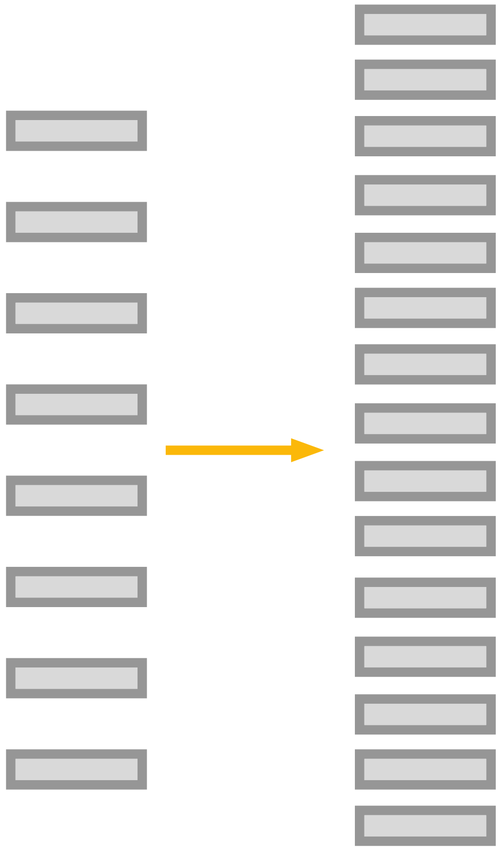
mapping



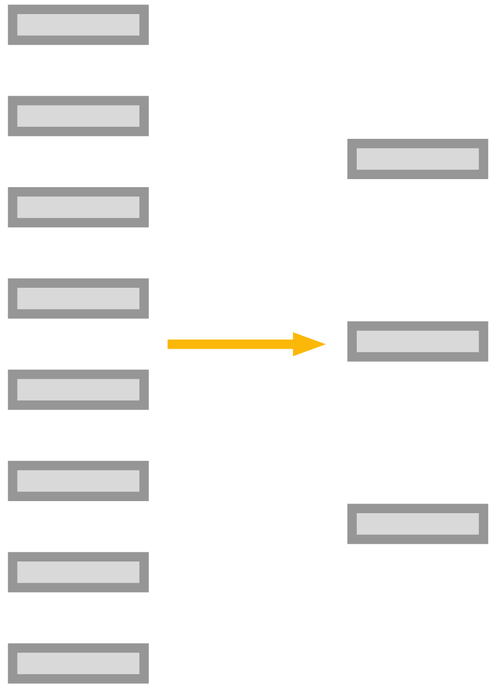
generating



mapping



generating



aggregating

```
spark_session.read.table("web.access_log")\
                .limit(3).toPandas()
```

```
spark_session.read.table("web.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...

```
spark_session.read.table("web.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...

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

```
import pyspark.sql.functions as f
```

```
import pyspark.sql.functions as f
```

```
access_log.select("user_agent",
                  f.length("user_agent"))\
          .limit(5).toPandas()
```

```
import pyspark.sql.functions as f
```

```
access_log.select("user_agent",
                  f.length("user_agent"))\
          .limit(5).toPandas()
```

	user_agent	length (user_agent)
0	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4)...	120
1	Mozilla/5.0 (Windows NT 5.1; U; de; rv:1.9.1.6...	88
2	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT...	153
3	Mozilla/5.0 (Linux; Android 4.4.2; nb-no; SAMS...	164
4	Mozilla/5.0 (Linux; Android 4.4.2; nb-no; SAMS...	164

```
access_log.select("user_agent",
                  f.length("user_agent").
                  alias("len"))\
          .limit(5).toPandas()
```

```
access_log.select("user_agent",
                  f.length("user_agent").
                  alias("len"))\
          .limit(5).toPandas()
```

	user_agent	length
0	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4)...	120
1	Mozilla/5.0 (Windows NT 5.1; U; de; rv:1.9.1.6...	88
2	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT...	153
3	Mozilla/5.0 (Linux; Android 4.4.2; nb-no; SAMS...	164
4	Mozilla/5.0 (Linux; Android 4.4.2; nb-no; SAMS...	164

```
access_log.select("url").limit(5).toPandas()
```

```
access_log.select("url").limit(5).toPandas()
```

	url
0	/id53821
1	/id33929
2	/id35754
3	/id78231
4	/id39395

```
access_log.select("url",
                  f.concat("http://vk.com",
                           "url")
                  ).limit(5).toPandas()
```

AnalysisException Traceback (most recent call last)
<ipython-input-98-601ab50e5a62> in <module>()
1 access_log.select("url",
2 f.concat("http://vk.com",
--> 3 "url")
4).limit(5).toPandas()

cannot resolve '`http://vk.com`' given input columns: [response_length, ip, time, url, http_code, user_agent];;

```
access_log.select("url",
                  f.concat(f.lit("http://vk.com"),
                           access_log_df.url)
                  ).limit(5).toPandas()
```

```
access_log.select("url",
                  f.concat(f.lit("http://vk.com"),
                           access_log_df.url)
                  ).limit(5).toPandas()
```

	url	concat (http://vk.com, url)
0	/id53821	http://vk.com/id53821
1	/id33929	http://vk.com/id33929
2	/id35754	http://vk.com/id35754
3	/id78231	http://vk.com/id78231
4	/id39395	http://vk.com/id39395


Search
Fedor Dostoevski
Profile hidden
VK © 2017
about
terms
developers

```
access_log.select("user_agent")\
          .limit(5).toPandas()
```

```
access_log.select("user_agent")\
          .limit(5).toPandas()
```

	user_agent
0	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4)...
1	Mozilla/5.0 (Windows NT 5.1; U; de; rv:1.9.1.6...
2	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT...
3	Mozilla/5.0 (Linux; Android 4.4.2; nb-no; SAMS...
4	Mozilla/5.0 (Linux; Android 4.4.2; nb-no; SAMS...

```
access_log.select("user_agent")\
          .select("user_agent",
                  f.split("user_agent"," ")
                    .alias("list")
                  )\
          .limit(5).toPandas()
```

```
access_log.select("user_agent")\
          .select("user_agent",
                  f.split("user_agent"," ")
                    .alias("list")
                  )\
          .limit(5).toPandas()
```

	user_agent	list
0	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4)...	[Mozilla/5.0 (Macintosh; Intel...
1	Mozilla/5.0 (Windows NT 5.1; U; de; rv:1.9.1.6...	[Mozilla/5.0 (Windows NT 5.1...
2	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT...	[Mozilla/4.0 (compatible; MSIE..
3	Mozilla/5.0 (Linux; Android 4.4.2; nb-no; SAMS...	[Mozilla/5.0 (Linux; Android...
4	Mozilla/5.0 (Linux; Android 4.4.2; nb-no; SAMS...	[Mozilla/5.0 (Linux; Android...

```
access_log.select("user_agent")\
             .select("user_agent",
                     f.split("user_agent"," ")
                       .alias("list")
                     )\
             .select("user_agent",
                       f.col("list")[0],
                       f.col("list")[1],
                       f.col("list")[2]
                     )\
             .limit(4).toPandas()
```

```
access_log.select("user_agent")\
             .select("user_agent",
                     f.split("user_agent"," ")
                       .alias("list")
                     )\
             .select("user_agent",
                       f.col("list")[0],
                       f.col("list")[1],
                       f.col("list")[2]
                     )\
             .limit(4).toPandas()
```

	user_agent	list[0]	list[1]	list[2]
0	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4)...	Mozilla/5.0	(Macintosh	Intel
1	Mozilla/5.0 (Windows NT 5.1; U; de; rv:1.9.1.6...	Mozilla/5.0	(Windows	NT
2	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT...	Mozilla/4.0	(compatible	MSIE
3	Mozilla/5.0 (Linux; Android 4.4.2; nb-no; SAMS...	Mozilla/5.0	(Linux	Android

```
access_log.select("user_agent")\
             .select("user_agent",
                     f.split("user_agent"," ")
                       .alias("list")
                     )\
             .select("user_agent",
                       f.explode("list"),
                     )\
             .limit(5).toPandas()
```

```
access_log.select("user_agent")\
            .select("user_agent",
                    f.split("user_agent"," ")
                      .alias("list")
                    )\
            .select("user_agent",
                      f.explode("list"),
                    )\
            .limit(5).toPandas()
```

	user_agent	col
0	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4)...	Mozilla/5.0
1	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4)...	(Macintosh;
2	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4)...	Intel
3	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4)...	Mac
4	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4)...	OS

```
access_log.select("user_agent")\
             .select("user_agent",
                     f.split("user_agent"," ")
                       .alias("list")
                     )\
             .select("user_agent",
                       f.explode("list"),
                     )\
             .where(f.col("col") == "Android")\
             .limit(3).toPandas()
```

```
access_log.select("user_agent")\
             .select("user_agent",
                     f.split("user_agent"," ")
                       .alias("list")
                     )\
             .select("user_agent",
                       f.explode("list"),
                     )\
             .where(f.col("col") == "Android")\
             .limit(3).toPandas()
```

	user_agent	col
0	Mozilla/5.0 (Linux; Android 4.4.2; nb-no; SAMS...	Android
1	Mozilla/5.0 (Linux; Android 4.4.2; nb-no; SAMS...	Android
2	Mozilla/5.0 (Linux; Android 4.4.1; Caesar Buil...	Android

```
access_log.select("user_agent")\
             .select("user_agent",
                     f.split("user_agent"," ")
                       .alias("list")
                     )\
             .select("user_agent",
                       f.explode("list"),
                     )\
             .where(f.col("col") == "Android")\
             .limit(3).toPandas()
```

	user_agent	col
0	Mozilla/5.0 (Linux; Android 4.4.2; nb-no; SAMS...	Android
1	Mozilla/5.0 (Linux; Android 4.4.2; nb-no; SAMS...	Android
2	Mozilla/5.0 (Linux; Android 4.4.1; Caesar Buil...	Android

```
access_log.select("user_agent")\
             .select("user_agent",
                     f.split("user_agent"," ")
                       .alias("list")
                     )\
             .select("user_agent",
                       f.explode("list"),
                     )\
             .where(f.col("col") == "Android")\
             .limit(3).toPandas()
```

	user_agent	col
0	Mozilla/5.0 (Linux; Android 4.4.2; nb-no; SAMS...	Android
1	Mozilla/5.0 (Linux; Android 4.4.2; nb-no; SAMS...	Android
2	Mozilla/5.0 (Linux; Android 4.4.1; Caesar Buil...	Android

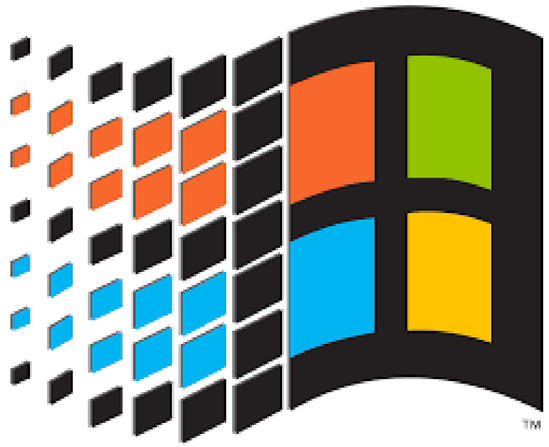
```
access_log.select("user_agent",
                  f.when(access_log.user_agent.like("%Android%"),
                         "Android")
                    .otherwise("Other")\
                    .alias("OS")
                ).limit(5).toPandas()
```

```
access_log.select("user_agent",
                  f.when(access_log.user_agent.like("%Android%"),
                         "Android")
                    .otherwise("Other")\
                    .alias("OS")
                ).limit(5).toPandas()
```

	user_agent	OS
0	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4)...	Other
1	Mozilla/5.0 (Windows NT 5.1; U; de; rv:1.9.1.6...	Other
2	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT...	Other
3	Mozilla/5.0 (Linux; Android 4.4.2; nb-no; SAMS...	Android
4	Mozilla/5.0 (Linux; Android 4.4.2; nb-no; SAMS...	Android

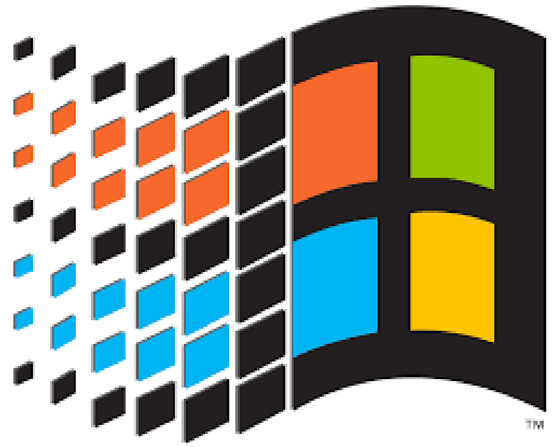
```
access_log.select("user_agent",
                  f.when(access_log.user_agent.like("%Android%"),
                         "Android")
                    .when(access_log.user_agent.like("%Windows%"),
                         "Windows")
                    .otherwise("Other")\
                    .alias("OS")
                ).limit(5).toPandas()
```

```
access_log.select("user_agent",
                  f.when(access_log.user_agent.like("%Android%"),
                         "Android")
                    .when(access_log.user_agent.like("%Windows%"),
                         "Windows")
                    .otherwise("Other")\
                    .alias("OS")
                ).limit(5).toPandas()
```



```
access_log.select("user_agent",
                  f.when(access_log.user_agent.like("%Android%"),
                         "Android")
                   .when(access_log.user_agent.like("%Windows%"),
                         "Windows")
                   .otherwise("Other")\
                   .alias("OS")
                ).limit(5).toPandas()
```

	user_agent	OS
0	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4)...	Other
1	Mozilla/5.0 (Windows NT 5.1; U; de; rv:1.9.1.6...	Windows
2	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT...	Windows
3	Mozilla/5.0 (Linux; Android 4.4.2; nb-no; SAMS...	Android
4	Mozilla/5.0 (Linux; Android 4.4.2; nb-no; SAMS...	Android



You have learned:

- types of SQL functions
- how to apply them to the DataFrame columns