Projection and Filtering

```
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...

```
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		
0	200	109.106.133.8	21546	12/Dec /2015:01:31:4 +0400		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		
0	200	109.106.133.8	21546	12/Dec /2015:01:31:4 +0400		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...

Projection

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

	http_code	ip	response_length	time		
0	200	109.106.133.8	21546	12/Dec /2015:01:31:4 +0400		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...

Filtering

Projection

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

```
spark_session.sql("""
    select ip, url
    from web.access_log
""").limit(3).toPandas()
```

```
spark_session.sql("""
    select ip, url
    from web.access_log
""").limit(3).toPandas()
```

	ip	url
0	109.106.133.8	/id53821
1	46.31.82.254	/id33929
2	193.124.254.46	/id35754

```
access_log_df.select("ip","url")\
             .limit(3).toPandas()
```

	ip	url
0	109.106.133.8	/id53821
1	46.31.82.254	/id33929
2	193.124.254.46	/id35754

```
access_log_df.select(access_log_df.ip,
                     access_log_df.url)\
             .limit(3).toPandas()
```

	ip	url
0	109.106.133.8	/id53821
1	46.31.82.254	/id33929
2	193.124.254.46	/id35754

```
access_log_df.select(access_log_df.ip,
                     access_log_df.url)\
             .limit(3).toPandas()
```

	ip	url
0	109.106.133.8	/id53821
1	46.31.82.254	/id33929
2	193.124.254.46	/id35754

```
access_log_df.select(access_log_df.ip,
                     access_log_df.url.alias("url_part"))\
             .limit(3).toPandas()
```

	ip	url
0	109.106.133.8	/id53821
1	46.31.82.254	/id33929
2	193.124.254.46	/id35754

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

```
access_log_df.select(f.col("ip"),
                     f.col("url").alias("url_part"))\
             .limit(3).toPandas()
```

	ip	url
0	109.106.133.8	/id53821
1	46.31.82.254	/id33929
2	193.124.254.46	/id35754

Field 1	Field 2	Unknown	b2a1rev_cpm

```
spark_session.sql("""
  select *
  from web.access_log
  where http_code<>'200'
""").limit(3).toPandas()
```

```
spark_session.sql("""
  select *
  from web.access_log
  where http_code<>'200'
""").limit(3).toPandas()
```

	http_code	ip	response_length	time	url	user_agent
0	404	91.206.117.71	0	12/Dec /2015:01:32:04 +0400	/favicon.ico	Mozilla/5.0 (X11; Linux x86_64; rv:31.0) Gecko...
1	404	23.39.172.114	0	12/Dec /2015:01:32:05 +0400	/favicon.ico	Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKi...
2	404	176.120.130.254	0	12/Dec /2015:01:32:06 +0400	/favicon.ico	Mozilla/5.0 (Linux; U; Android 4.2.2; de-at; H...

Where

Filter

Where

Filter

```
access_log_df.where("http_code <> '200'").limit(3).toPandas()
```

Where

Filter

```
access_log_df.where("http_code <> '200'").limit(3).toPandas()
```

	http_code	ip	response_length	time	url	user_agent
0	404	91.206.117.71	0	12/Dec /2015:01:32:04 +0400	/favicon.ico	Mozilla/5.0 (X11; Linux x86_64; rv:31.0) Gecko...
1	404	23.39.172.114	0	12/Dec /2015:01:32:05 +0400	/favicon.ico	Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKi...
2	404	176.120.130.254	0	12/Dec /2015:01:32:06 +0400	/favicon.ico	Mozilla/5.0 (Linux; U; Android 4.2.2; de-at; H...

```
access_log_df.where(access_log_df.http_code <> '200')\
                .limit(3).toPandas()
```

```
access_log_df.where(access_log_df.http_code <> '200')\
                .limit(3).toPandas()
```

	http_code	ip	response_length	time	url	user_agent
0	404	91.206.117.71	0	12/Dec /2015:01:32:04 +0400	/favicon.ico	Mozilla/5.0 (X11; Linux x86_64; rv:31.0) Gecko...
1	404	23.39.172.114	0	12/Dec /2015:01:32:05 +0400	/favicon.ico	Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKi...
2	404	176.120.130.254	0	12/Dec /2015:01:32:06 +0400	/favicon.ico	Mozilla/5.0 (Linux; U; Android 4.2.2; de-at; H...

```
spark_session.sql("""
  select *
  from web.access_log
  where http_code<>'200'
  and user_agent like '%Android%'
""").limit(3).toPandas()
```

```
spark_session.sql("""
  select *
  from web.access_log
  where http_code<>'200'
  and user_agent like '%Android%'
""").limit(3).toPandas()
```

	http_code	ip	response_length	time	url	user_agent
0	404	176.120.130.254	0	12/Dec /2015:01:32:06 +0400	/favicon.ico	Mozilla/5.0 (Linux; U; Android 4.2.2; de-at; H...
1	404	93.188.131.176	0	12/Dec /2015:01:32:07 +0400	/favicon.ico	Mozilla/5.0 (Linux; U; Android 4.2.2; de-de; L...
2	404	87.245.244.151	0	12/Dec /2015:01:32:08 +0400	/favicon.ico	Mozilla/5.0 (Linux; Android 5.1.1; D6603 Build...

```
access_log_df.where(
     (access_log_df.http_code <> '200') &
     (access_log_df.user_agent.like('%Android%')))\
     .limit(3).toPandas()
```

```
access_log_df.where(
    (access_log_df.http_code <> '200') &
    (access_log_df.user_agent.like('%Android%')))\
    .limit(3).toPandas()
```

	http_code	ip	response_length	time	url	user_agent
0	404	176.120.130.254	0	12/Dec /2015:01:32:06 +0400	/favicon.ico	Mozilla/5.0 (Linux; U; Android 4.2.2; de-at; H...
1	404	93.188.131.176	0	12/Dec /2015:01:32:07 +0400	/favicon.ico	Mozilla/5.0 (Linux; U; Android 4.2.2; de-de; L...
2	404	87.245.244.151	0	12/Dec /2015:01:32:08 +0400	/favicon.ico	Mozilla/5.0 (Linux; Android 5.1.1; D6603 Build...

```
access_log_df.[["url","ip"]]\
                .limit(3).toPandas()
```

	url	id
0	/id53821	109.106.133.8
1	/id33929	46.31.82.254
2	/id35754	193.124.254.46

```
access_log_df[access_log_df.http_code <> '200']\
                    .limit(3).toPandas()
```

	http_code	ip	response_length	time	url	user_agent
0	404	91.206.117.71	0	12/Dec /2015:01:32:04 +0400	/favicon.ico	Mozilla/5.0 (X11; Linux x86_64; rv:31.0) Gecko...
1	404	23.39.172.114	0	12/Dec /2015:01:32:05 +0400	/favicon.ico	Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKi...
2	404	176.120.130.254	0	12/Dec /2015:01:32:06 +0400	/favicon.ico	Mozilla/5.0 (Linux; U; Android 4.2.2; de-at; H...

What have you learned:

- How to make projections in dataframe API
- How to make filtering