

ASSIGNMENT NO. 13

```
Activities Terminal Apr 16 15:26 student@student: ~/TB62

(base) student@student: $ cd TB62
(base) student@student: ~/TB62$ spark-shell
24/04/16 15:24:10 WARN Utils: Your hostname, student resolves to a loopback address: 127.0.0.1; using 10.11.5.72 instead (on interface enp3s0)
24/04/16 15:24:10 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/04/16 15:24:27 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Spark context Web UI available at http://10.11.5.72:4040
Spark context available as 'sc' (master = local[*], app id = local-1713261269083).
Spark session available as 'spark'.
Welcome to

      ____
     / ___/
    / __/   version 3.5.1
   /___/

Using Scala version 2.12.18 (OpenJDK 64-Bit Server VM, Java 11.0.22)
Type in expressions to have them evaluated.
Type :help for more information.

scala> :load WebLog_Processing.scala
Loading WebLog_Processing.scala...
import org.apache.log4j.{Level, Logger}
import org.apache.spark.sql.{Column, SparkSession}
import org.apache.spark.sql.functions.{regexp_extract, sum, col, to_date, udf, to_timestamp, desc, dayofyear, year}
24/04/16 15:24:59 WARN SparkSession: Using an existing Spark session; only runtime SQL configurations will take effect.
spark: org.apache.spark.sql.SparkSession = org.apache.spark.sql.SparkSession@611f774a
base_df: org.apache.spark.sql.DataFrame = [value: string]
root
|-- value: string (nullable = true)

import spark.implicits._
base_df: org.apache.spark.sql.DataFrame = [value: string]
root
|-- value: string (nullable = true)

+-----+
|value|
+-----+
|IP,Time,URL,Staus|
|10.128.2.1|[29/Nov/2017:06:58:55,GET /login.php HTTP/1.1,200|
|10.128.2.1|[29/Nov/2017:06:59:02,POST /process.php HTTP/1.1,302|
+-----+
only showing top 3 rows
```

```
Activities Terminal Apr 16 15:26 student@student: ~/TB62

|IP,Time,URL,Staus|
|10.128.2.1|[29/Nov/2017:06:58:55,GET /login.php HTTP/1.1,200|
|10.128.2.1|[29/Nov/2017:06:59:02,POST /process.php HTTP/1.1,302|
+-----+
only showing top 3 rows

parsed_df: org.apache.spark.sql.DataFrame = [host: string, timestamp: string ... 2 more fields]
+-----+
|host|timestamp|path|status|
+-----+
|IP| | |NULL|
|10.128.2.1|29/Nov/2017:06:58:55|/login.php|200|
|10.128.2.1|29/Nov/2017:06:59:02|/process.php|302|
|10.128.2.1|29/Nov/2017:06:59:03|/home.php|200|
|10.131.2.1|29/Nov/2017:06:59:04|/js/vendor/moment.min.js|200|
+-----+
only showing top 5 rows

root
|-- host: string (nullable = true)
|-- timestamp: string (nullable = true)
|-- path: string (nullable = true)
|-- status: integer (nullable = true)

Number of bad row in the initial dataset : 0
bad_row_df: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [host: string, timestamp: string ... 2 more fields]
Number of bad rows : 219
count_null: (col_name: org.apache.spark.sql.Column)org.apache.spark.sql.Column
t: Array[org.apache.spark.sql.Column] = Array(sum(CAST((host IS NULL) AS INT)) AS host, sum(CAST((timestamp IS NULL) AS INT)) AS timestamp, sum(CAST((path IS NULL) AS INT)) AS path, s
um(CAST((status IS NULL) AS INT)) AS status)
+-----+
|host|timestamp|path|status|
+-----+
|0|0|0|219|
+-----+

bad_status_df: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [bad_status: string]
Number of bad rows : 219
+-----+
|bad_status|
+-----+
|IP,Time,URL,Staus|
|chmod:,cannot,'a....|
|chmod:,cannot,'er...|
|rm:,cannot,'*o!;,No|
|rm:,cannot,'a,out...|
+-----+
```

```

Activities Terminal Apr 16 15:26 student@student: ~/TB62

|rm: cannot, 'a.out...|
+-----+
only showing top 5 rows

cleaned_df: org.apache.spark.sql.DataFrame = [host: string, timestamp: string ... 2 more fields]
The count of null value : 0
Before : 16008 | After : 15789
+-----+
|to_date(timestamp)|
+-----+
| NULL |
| NULL |
+-----+
only showing top 2 rows

month_map: scala.collection.immutable.Map[String,Int] = Map(Nov -> 11, Jul -> 7, Mar -> 3, Jan -> 1, Oct -> 10, Dec -> 12, Feb -> 2, May -> 5, Apr -> 4, Aug -> 8, Sep -> 9, Jun -> 6)
parse_clf_time: (s: String)String
toTimestamp: org.apache.spark.sql.expressions.UserDefinedFunction = SparkUserDefinedFunction($Lambda$4635/0x00000000841931040@40b1bc9a,StringType,List(Some(class[value@0]: string))),So
ne(class[value@0]: string)),None,true,true)
log_df: org.apache.spark.sql.DataFrame = [host: string, path: string ... 2 more fields]
root
|-- host: string (nullable = true)
|-- path: string (nullable = true)
|-- status: integer (nullable = true)
|-- time: timestamp (nullable = true)
+-----+
| host | path | status | time |
+-----+
| 10.128.2.1 | /login.php | 200 | 2017-11-29 06:58:55 |
| 10.128.2.1 | /process.php | 302 | 2017-11-29 06:59:02 |
+-----+
only showing top 2 rows

res29: logs_df.type = [host: string, path: string ... 2 more fields]
+-----+
|summary| status|
+-----+
| count | 15789 |
| mean | 230.19469250744189 |
| stddev | 50.05853522906924 |
| min | 200 |
| max | 404 |
+-----+
+-----+

```

```

Activities Terminal Apr 16 15:26 student@student: ~/TB62

+-----+
|status|count|
+-----+
| 200 | 11330 |
| 206 | 52 |
| 302 | 3498 |
| 304 | 658 |
| 404 | 251 |
+-----+

+-----+
| host | count |
+-----+
| 10.131.2.1 | 1626 |
| 10.128.2.1 | 4257 |
| 10.130.2.1 | 4056 |
| 10.131.0.1 | 4198 |
| 10.129.2.1 | 1652 |
+-----+

+-----+
| path | count |
+-----+
| /login.php | 3298 |
| /hone.php | 2653 |
| /js/vendor/modern... | 1417 |
| / | 862 |
| /contestproblem.p... | 467 |
| /css/normalize.css | 408 |
| /css/bootstrap.mi... | 404 |
| /css/font-awesome... | 399 |
| /css/style.css | 395 |
| /css/main.css | 394 |
| /js/vendor/jquery... | 387 |
| /bootstrap-3.3.7/... | 382 |
| /process.php | 317 |
| /contest.php | 249 |
| /archive.php | 246 |
| /fonts/fontawesome... | 245 |
| /robots.txt | 224 |
| /img/ruet.png | 213 |
| /bootstrap-3.3.7/... | 191 |
| /js/vendor/moment... | 173 |
+-----+
only showing top 20 rows

```

```
Activities Terminal Apr 16 15:26 student@student: ~/TB62

+-----+
| path | count |
+-----+
| /login.php | 3298 |
| /home.php | 2653 |
| /js/vendor/modern... | 1417 |
| / | 862 |
| /contestproblem.p... | 467 |
| /css/normalize.css | 488 |
| /css/bootstrap.mi... | 484 |
| /css/font-awesome... | 399 |
| /css/style.css | 395 |
| /css/main.css | 394 |
+-----+
only showing top 10 rows

+-----+
| path | count |
+-----+
| /home.php | 2167 |
| / | 741 |
| /process.php | 317 |
| /robots.txt | 224 |
| /action.php | 83 |
| /contestproblem.p... | 74 |
| /js/vendor/jquery... | 73 |
| /css/bootstrap.mi... | 72 |
| /js/vendor/modern... | 72 |
| /css/main.css | 68 |
+-----+
only showing top 10 rows

unique_host_count: Long = 5
Unique hosts: 5
daily_hosts_df: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [day: int, year: int ... 1 more field]

+-----+
| day | year | count |
+-----+
| 311 | 2017 | 1 |
| 312 | 2017 | 5 |
| 313 | 2017 | 5 |
| 314 | 2017 | 5 |
| 315 | 2017 | 5 |
+-----+
only showing top 5 rows
```

```
Activities Terminal Apr 16 15:27 student@student: ~/TB62

only showing top 5 rows

total_req_per_day_df: org.apache.spark.sql.DataFrame = [day: int, year: int ... 1 more field]
avg_daily_request_per_host_df: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [day: int, year: int ... 1 more field]

+-----+
| day | year | avg_req_per_host_per_day |
+-----+
| 335 | 2017 | 93.6 |
| 327 | 2017 | 76.8 |
| 60 | 2018 | 10.333333333333334 |
| 350 | 2017 | 51.666666666666664 |
| 46 | 2018 | 6.666666666666667 |
+-----+
only showing top 5 rows

not_found_df: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [host: string, path: string ... 2 more fields]
found 251 404 Urls

+-----+
| path |
+-----+
| /css/bootstrap.min.css.map |
| /robots.txt |
| /djs/vendor/bootstrap-datetimepicker.js |
| /favicon.ico |
+-----+

+-----+
| path | count |
+-----+
| /css/bootstrap.min.css.map | 1 |
| /djs/vendor/bootstrap-datetimepicker.js | 7 |
| /favicon.ico | 19 |
| /robots.txt | 224 |
+-----+

+-----+
| path | collect_list(host) | count(status) |
+-----+
| /css/bootstrap.mi... | [10.130.2.1] | 1 |
| /djs/vendor/boots... | [10.131.0.1, 10.1... | 7 |
| /favicon.ico | [10.128.2.1, 10.1... | 19 |
| /robots.txt | [10.131.0.1, 10.1... | 224 |
+-----+

+-----+
| path | collect_set(host) | count(status) |
+-----+
```

```
Activities Terminal Apr 16 15:27 student@student: ~/TB62

| path| collect_list(host)|count(status)|
+-----+-----+
|/css/bootstrap.mi...| [10.130.2.1]| 1|
|/djs/vendor/boots...| [10.131.0.1, 10.1...| 7|
|/favicon.ico|[10.128.2.1, 10.1...| 19|
|/robots.txt|[10.131.0.1, 10.1...| 224|
+-----+-----+

| path| collect_set(host)|count(status)|
+-----+-----+
|/css/bootstrap.mi...| [10.130.2.1]| 1|
|/djs/vendor/boots...| [10.130.2.1, 10.1...| 7|
|/favicon.ico|[10.130.2.1, 10.1...| 19|
|/robots.txt|[10.130.2.1, 10.1...| 224|
+-----+-----+

+-----+-----+
|host| count|
+-----+-----+
|10.128.2.1|67|
|10.131.0.1|61|
|10.130.2.1|52|
|10.129.2.1|41|
|10.131.2.1|30|
+-----+-----+

errors_by_date_pair_df: org.apache.spark.sql.DataFrame = [day: int, year: int ... 1 more field]
+-----+-----+
|day|year|count|
+-----+-----+
|312|2017| 8|
|313|2017| 10|
|314|2017| 6|
|315|2017| 12|
|316|2017| 6|
|317|2017| 10|
|318|2017| 10|
|319|2017| 8|
|320|2017| 10|
|321|2017| 5|
+-----+-----+
only showing top 10 rows

scala>
```