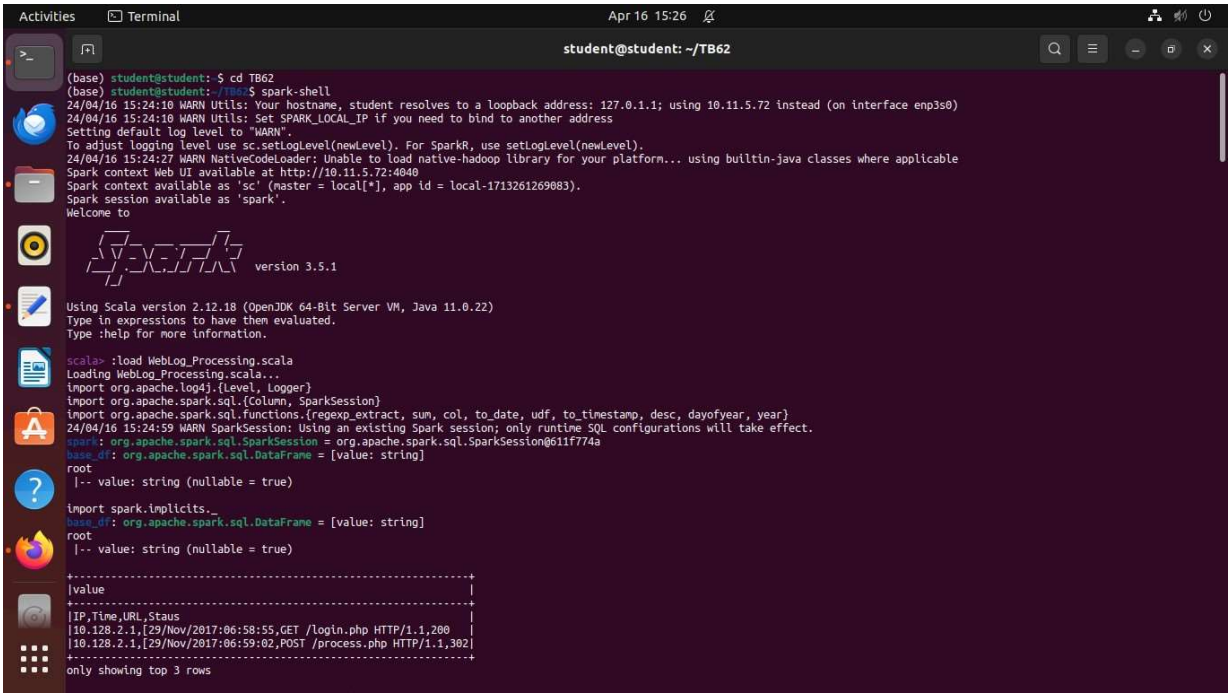
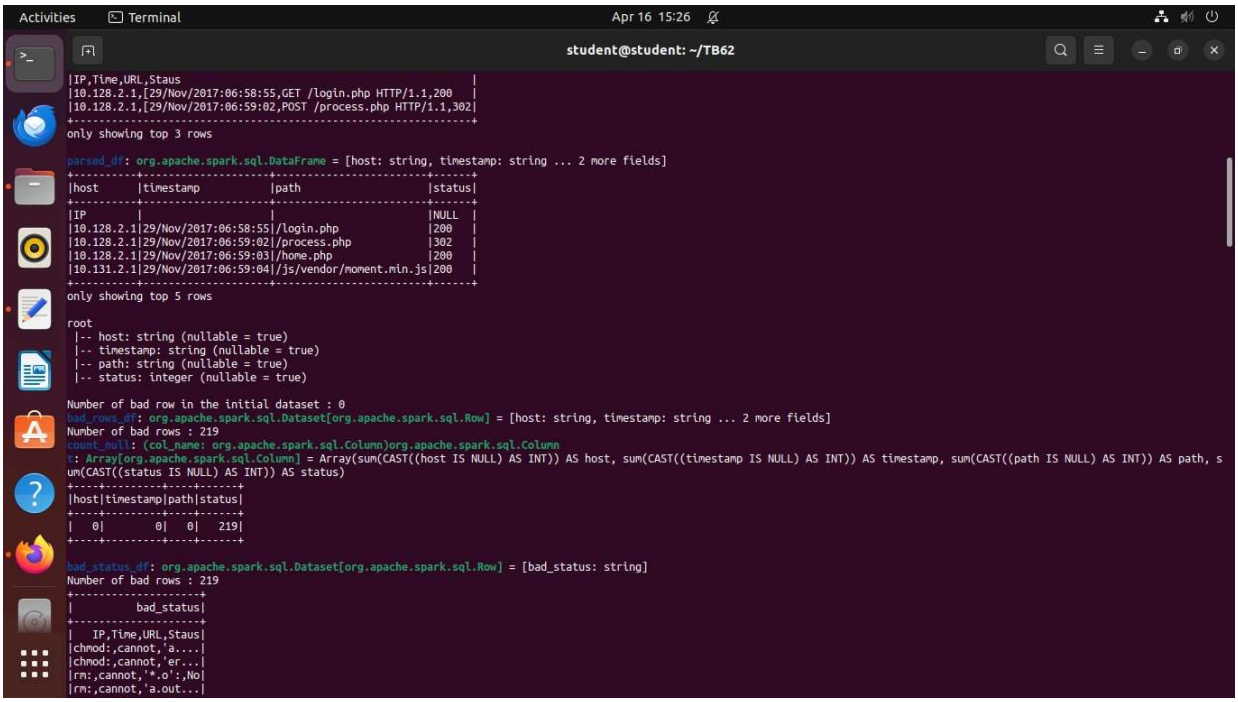
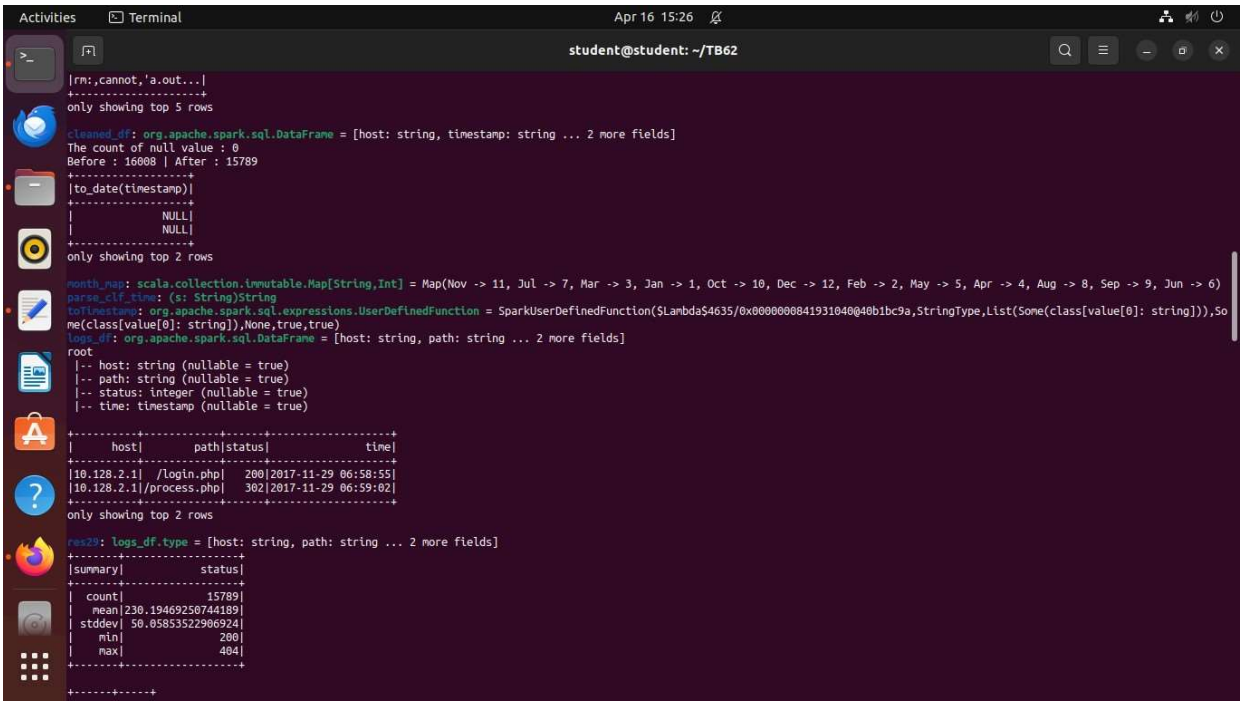
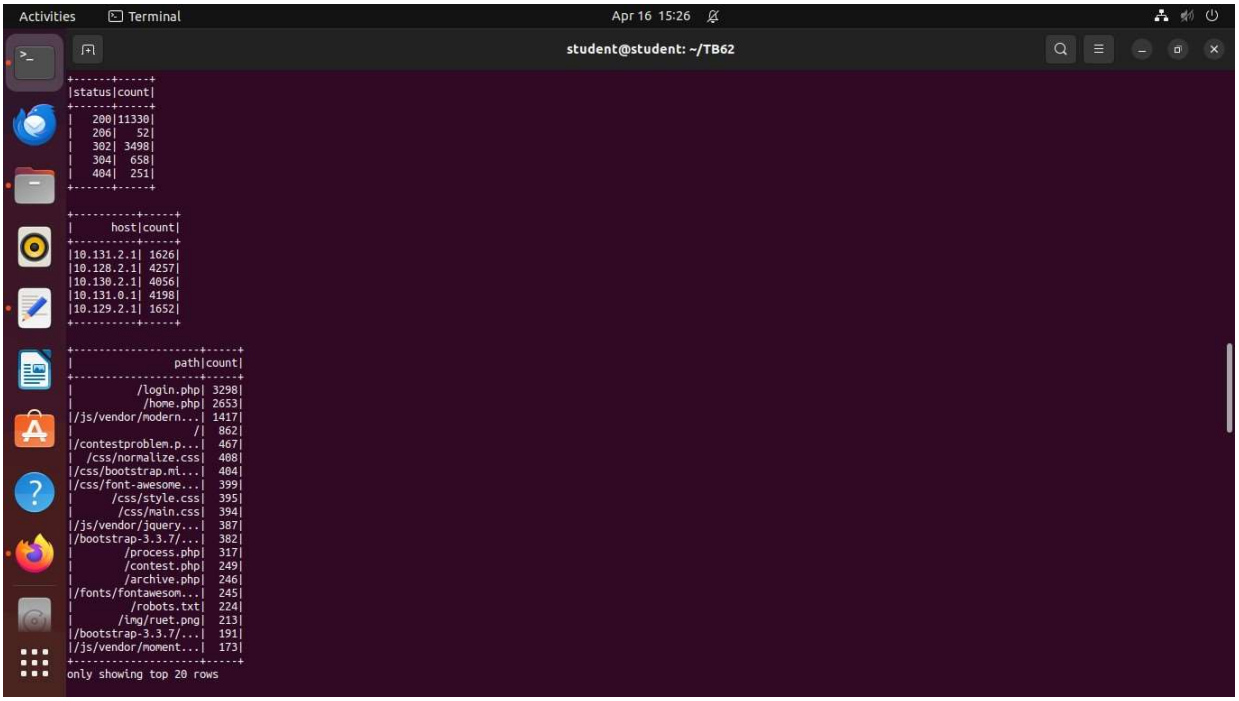
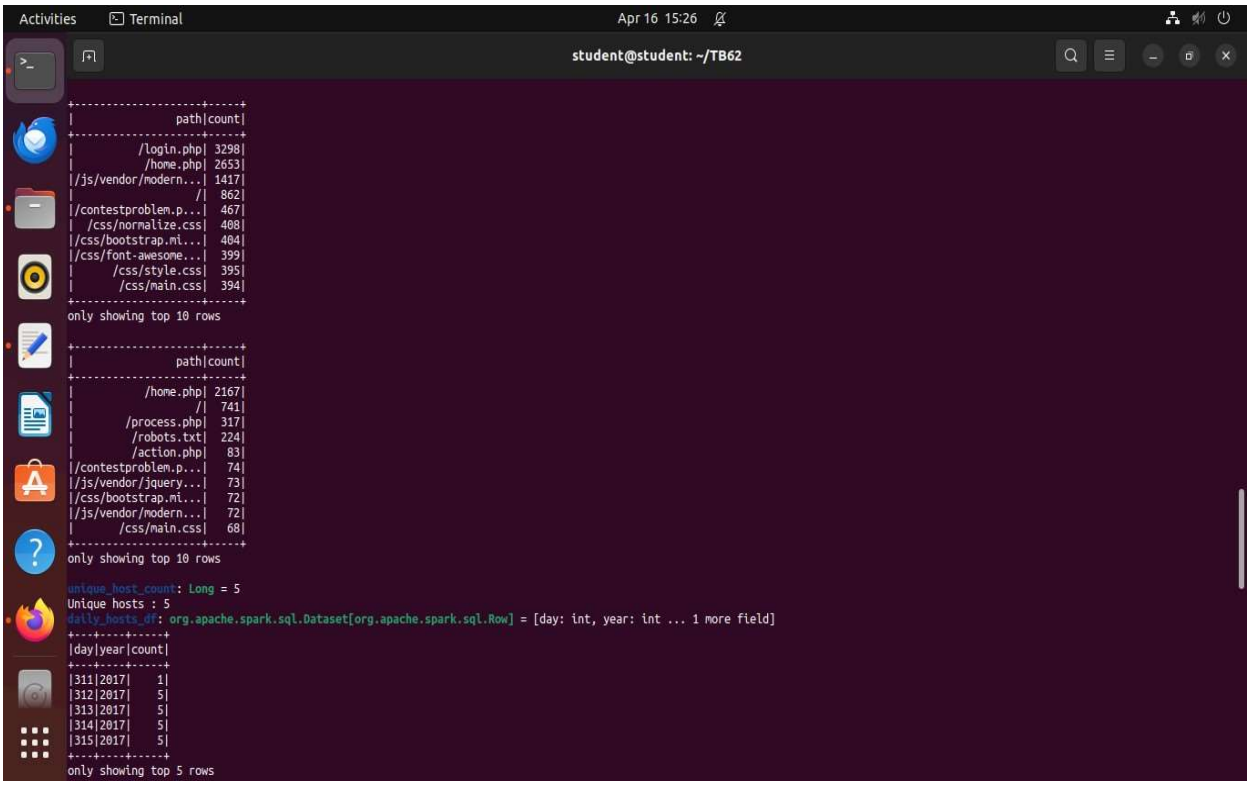
ASSIGNMENT NO. 13

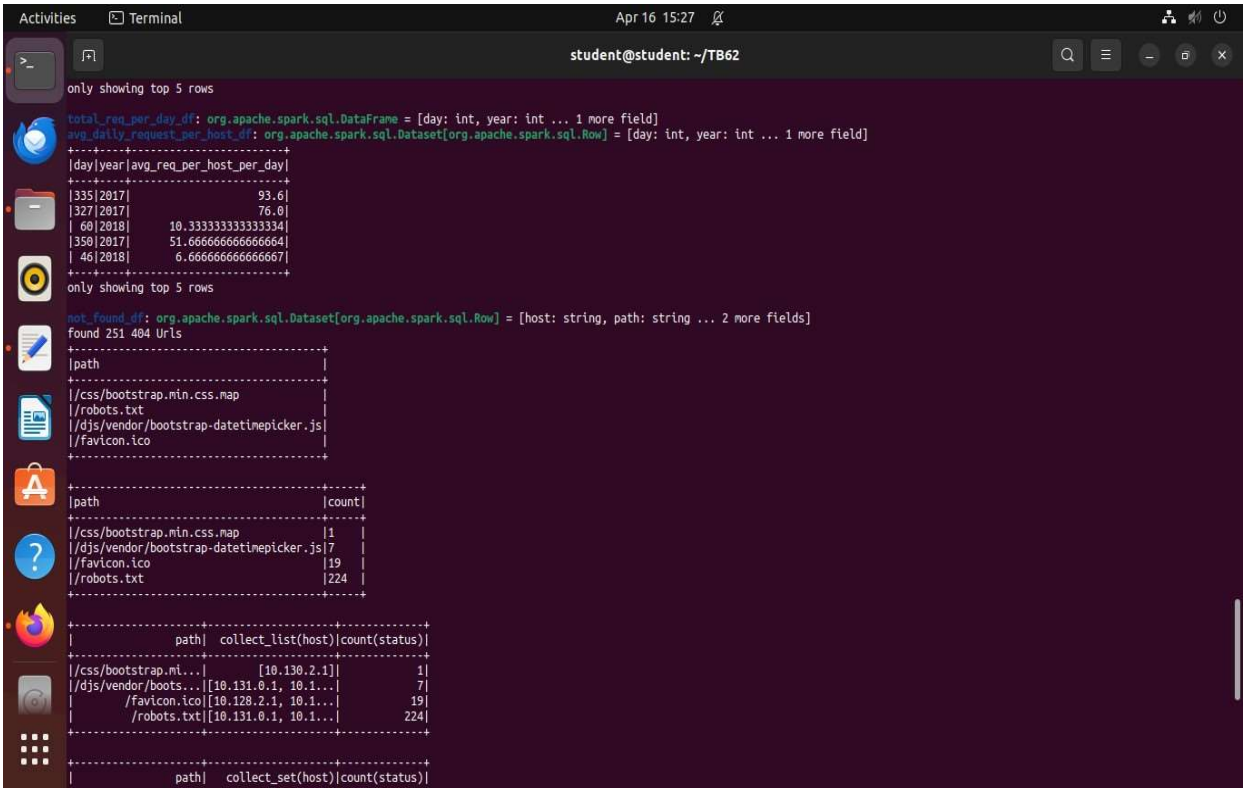


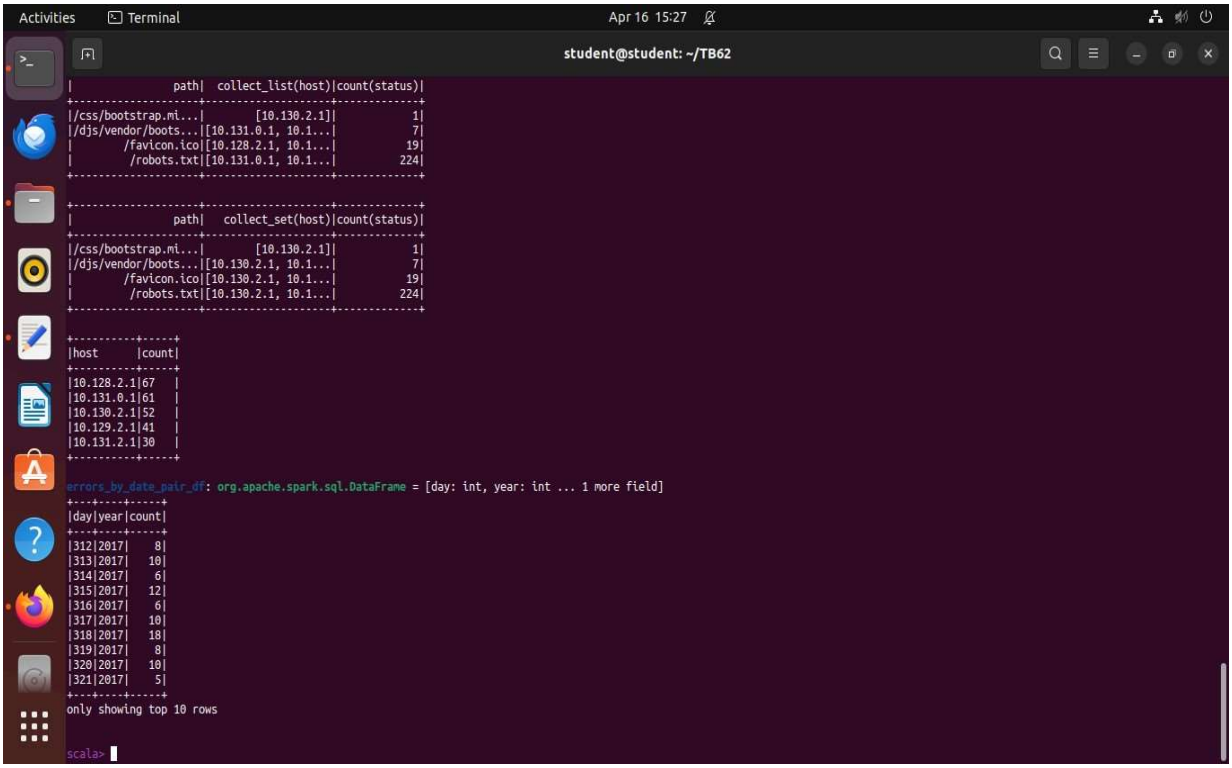



Activities
Terminal
Apr 16 15:26
student@student: ~/TB62
|rm:,cannot,'a.out...|
only showing top 5 rows
cleaned_df: org.apache.spark.sql.DataFrame = [host: string, timestamp: string ... 2 more fields]
The count of null value : 0
Before : 16008 | After : 15789
|to_date(timestamp)|
NULL|
NULL|
only showing 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$4635/0x0000000841931040@40b1bc9a,StringType,List(Some(class[value[0]: string])),Some(class[value[0]: string]),None,true,true)
logs_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
/home.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/fontawesom...	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| 408|
|/css/bootstrap.mi...| 404|
|/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	76.0
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	18
319	2017	8
320	2017	10
321	2017	5
only showing top 10 rows
scala>