

```
In [5]: from pyspark.sql import SparkSession
        from pyspark.sql.functions import col, count, avg, min, max, to_date, hour
        import re
```

```
In [6]: spark=SparkSession.builder.master("local[*]").appName("NASAWebLogEDA").getOrCreate()
```

```
WARNING: Using incubator modules: jdk.incubator.vector
Using Spark's default log4j profile: org/apache/spark/log4j2-defaults.properties
26/02/04 13:40:23 WARN Utils: Your hostname, pl2-HP-280-Pro-G6-Microtower-PC, resolves to a loopback address: 127.0.1.1; using 192.168.29.209 instead (on interface enp1s0)
26/02/04 13:40:23 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
Using Spark's default log4j profile: org/apache/spark/log4j2-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
26/02/04 13:40:24 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

```
In [12]: log_file_path="/home/pl2/a2_39/"
        logs_rdd=spark.sparkContext.textFile(log_file_path)
```

```
In [13]: log_pattern = r'^(\S+) (\S+) (\S+) \[(.*?)\] "(\S+) (\S+) (\S+)" (\d{3}) (\S+)'

        def parse_line(line):
            match = re.match(log_pattern, line)
            if match:
                host, _, _, datetime, method, endpoint, protocol, status, content_size =
                    if content_size == '-':
                        content_size = 0
                    return (host, datetime, method, endpoint, protocol, int(status), int(content_size))
            else:
                return None

        parsed_logs_rdd = logs_rdd.map(parse_line).filter(lambda x: x is not None)

        columns = ["host", "datetime", "method", "endpoint", "protocol", "status", "content_size"]
        logs_df = parsed_logs_rdd.toDF(columns)
```

```
In [14]: logs_df.select(
            avg("content_size").alias("avg_size"),
            min("content_size").alias("min_size"),
            max("content_size").alias("max_size")
        ).show()
```

```
[Stage 2:=====> (3 + 5) / 8]
```

```
+-----+-----+-----+
|      avg_size|min_size|max_size|
+-----+-----+-----+
|20385.749398729513|      0|6823936|
+-----+-----+-----+
```

```
In [15]: logs_df.groupBy("status").count().show()
```

[Stage 5:=====> (6 + 2) / 8]

```
+-----+-----+
|status|  count|
+-----+-----+
|   304| 132626|
|   404|  10711|
|   500|     62|
|   403|     54|
|   200|1696830|
|   302|  46541|
|   501|     14|
+-----+-----+
```

```
In [16]: logs_df.groupBy("host").count().orderBy(col("count").desc()).show(10)
```

[Stage 8:=====> (7 + 1) / 8]

```
+-----+-----+-----+
|          host|count|
+-----+-----+-----+
|piweba3y.prodigy.com|17572|
|piweba4y.prodigy.com|11591|
|piweba1y.prodigy.com| 9868|
|  alyssa.prodigy.com| 7852|
| siltb10.orl.mmc.com| 7573|
|piweba2y.prodigy.com| 5922|
|  edams.ksc.nasa.gov| 5428|
|      163.206.89.4| 4906|
|      news.ti.com| 4863|
|disarray.demon.co.uk| 4353|
+-----+-----+-----+
only showing top 10 rows
```

```
In [17]: logs_df.groupBy("endpoint").count().orderBy(col("count").desc()).show(20)
```

```
+-----+-----+
|          endpoint| count|
+-----+-----+
|/images/NASA-logo...|111086|
|/images/KSC-logos...| 89529|
|/images/MOSAIC-lo...| 60299|
|/images/USA-logos...| 59844|
|/images/WORLD-log...| 59324|
|/images/ksclogo-m...| 58615|
|/images/launch-lo...| 40841|
| /shuttle/countdown/| 40248|
|          /ksc.html| 40064|
|/images/ksclogosm...| 33555|
|          /| 32669|
|/history/apollo/i...| 31052|
|/shuttle/missions...| 24833|
|  /htbin/cdt_main.pl| 22601|
|/shuttle/countdow...| 22189|
|/shuttle/countdow...| 21977|
|/shuttle/countdow...| 20920|
|/images/launchmed...| 20788|
|/shuttle/missions...| 19831|
|/shuttle/missions...| 18135|
+-----+-----+
only showing top 20 rows
```

```
In [18]: logs_df.filter(col("status") >= 400).groupBy("endpoint").count().orderBy(col("co
```

```
[Stage 14:=====> (1 + 7) / 8]
```

```
[Stage 14:=====> (2 + 6) / 8]
```

```
+-----+-----+
|          endpoint|count|
+-----+-----+
|/pub/winvn/readme...| 667|
|/pub/winvn/releas...| 547|
|/history/apollo/a...| 286|
|/shuttle/resource...| 230|
|/history/apollo/a...| 230|
|/://spacelink.msf...| 215|
|/history/apollo/p...| 215|
|/images/crawlerwa...| 214|
|/history/apollo/s...| 183|
|/shuttle/resource...| 180|
+-----+-----+
only showing top 10 rows
```

```
[Stage 14:=====> (4 + 4) / 8]
```

```
In [19]: unique_hosts = logs_df.select("host").distinct().count()
print("Unique hosts:", unique_hosts)
```

```
[Stage 17:=====> (6 + 2) / 8]
```

```
Unique hosts: 81892
```

```
In [20]: logs_df = logs_df.withColumn("date", to_date(col("datetime"), "dd/MMM/yyyy:HH:mm
requests_per_host_per_day = logs_df.groupBy("host", "date").count()
```

```
requests_per_host_per_day.groupBy().agg(avg("count")).show()
```

```
[Stage 23:=====> (7 + 1) / 8]
+-----+
|      avg(count)|
+-----+
|14.247598767669446|
+-----+
```

```
In [21]: logs_df.filter(col("status") == 404).groupBy("endpoint").count().orderBy(col("count"))
```

```
[Stage 29:=====> (5 + 3) / 8]
+-----+-----+
|      endpoint|count|
+-----+-----+
|/pub/winvn/readme...| 667|
|/pub/winvn/releas...| 547|
|/history/apollo/a...| 286|
|/shuttle/resource...| 230|
|/history/apollo/a...| 230|
|/://spacelink.msf...| 215|
|/history/apollo/p...| 215|
|/images/crawlerwa...| 214|
|/history/apollo/s...| 183|
|/shuttle/resource...| 180|
|/shuttle/missions...| 175|
|/shuttle/missions...| 168|
|/elv/DELTA/uncons...| 163|
|/history/apollo/p...| 140|
|/shuttle/missions...| 107|
|/shuttle/resource...| 92|
|/procurement/proc...| 86|
|/history/apollo-1...| 73|
|/history/apollo/p...| 71|
|/shuttle/countdow...| 68|
+-----+-----+
only showing top 20 rows
```

```
In [22]: logs_df.filter(col("status") == 404).groupBy("host").count().orderBy(col("count"))
```

```
[Stage 32:=====> (2 + 6) / 8]
```

```

+-----+-----+
|          host |count|
+-----+-----+
|hoohoo.ncsa.uiuc.edu| 251|
|jbiagioni.npt.nuw...| 131|
|piweba3y.prodigy.com| 110|
|piweba1y.prodigy.com|  92|
|phaelon.ksc.nasa.gov|  64|
|www-d4.proxy.aol.com|  61|
|piweba4y.prodigy.com|  56|
|monarch.eng.buffa...|  56|
|  alyssa.prodigy.com|  54|
|          titan02f|  53|
|www-a2.proxy.aol.com|  52|
|www-b4.proxy.aol.com|  48|
|www-b6.proxy.aol.com|  44|
|www-b3.proxy.aol.com|  43|
|tearnest2.stpaul....|  42|
|www-b2.proxy.aol.com|  41|
|www-d1.proxy.aol.com|  41|
|www-d3.proxy.aol.com|  38|
|piweba2y.prodigy.com|  38|
|www-a1.proxy.aol.com|  38|

```

only showing top 20 rows

[Stage 32:=====>

(4 + 4) / 8]

In [23]: `errors_per_day = logs_df.filter(col("status") == 404).groupBy("date").count().or
errors_per_day.show()`

[Stage 35:=====>

(6 + 2) / 8]

```

+-----+-----+
|      date |count|
+-----+-----+
|1995-07-01|  158|
|1995-07-02|  310|
|1995-07-03|  370|
|1995-07-04|  409|
|1995-07-05|  414|
|1995-07-06|  576|
|1995-07-07|  604|
|1995-07-08|  420|
|1995-07-09|  312|
|1995-07-10|  359|
|1995-07-11|  478|
|1995-07-12|  453|
|1995-07-13|  547|
|1995-07-14|  405|
|1995-07-15|  322|
|1995-07-16|  216|
|1995-07-17|  381|
|1995-07-18|  379|
|1995-07-19|  665|
|1995-07-20|  446|

```

only showing top 20 rows

In [24]: `errors_per_day.orderBy(col("count").desc()).show(3)`

[Stage 38:===== > (7 + 1) / 8]

```
+-----+-----+
|      date|count|
+-----+-----+
|1995-07-19|  665|
|1995-07-07|  604|
|1995-07-06|  576|
+-----+-----+
```

only showing top 3 rows

In [25]: `logs_df = logs_df.withColumn("hour", hour(to_date(col("datetime"), "dd/MMM/yyyy:HH:mm:ss"), "dd/MMM/yyyy:HH:mm:ss"))
logs_df.filter(col("status") == 404).groupBy("hour").count().orderBy("hour").show(1)`

[Stage 41:===== > (7 + 1) / 8]

```
+-----+-----+
|hour|count|
+-----+-----+
|  0|10711|
+-----+-----+
```

In [26]: `spark.stop()`

In []: