hotlinks2

October 4, 2022

[1]:
```
import wmfdata
spark = wmfdata.spark.get_session(type='yarn-large')
import pyspark.sql.functions as F
```

```
You are using wmfdata v1.3.2, but v1.3.3 is available.

To update, run `pip install --upgrade git+https://github.com/wikimedia/wmfdata-
python.git@release --ignore-installed`.

To see the changes, refer to https://github.com/wikimedia/wmfdata-
python/blob/release/CHANGELOG.md
PySpark executors will use /usr/lib/anaconda-wmf/bin/python3.

PYSPARK_PYTHON=/usr/lib/anaconda-wmf/bin/python3

SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/spark2/jars/slf4j-log4j12-1.7.16.jar!
/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-reload4j-1.7.36.jar!
/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use
setLogLevel(newLevel).
22/10/04 19:07:36 WARN SparkConf: Note that spark.local.dir will be overridden
by the value set by the cluster manager (via SPARK_LOCAL_DIRS in
mesos/standalone/kubernetes and LOCAL_DIRS in YARN).
22/10/04 19:07:37 WARN Utils: Service 'sparkDriver' could not bind on port
12000. Attempting port 12001.
22/10/04 19:07:37 WARN Utils: Service 'sparkDriver' could not bind on port
12001. Attempting port 12002.
22/10/04 19:07:37 WARN Utils: Service 'sparkDriver' could not bind on port
12002. Attempting port 12003.
22/10/04 19:07:38 WARN Utils: Service 'SparkUI' could not bind on port 4040.
Attempting port 4041.
22/10/04 19:07:38 WARN Utils: Service 'SparkUI' could not bind on port 4041.
Attempting port 4042.
22/10/04 19:07:38 WARN Utils: Service 'SparkUI' could not bind on port 4042.
```

```
Attempting port 4043.
22/10/04 19:07:49 WARN Utils: Service
'org.apache.spark.network.netty.NettyBlockTransferService' could not bind on
port 13000. Attempting port 13001.
22/10/04 19:07:49 WARN Utils: Service
'org.apache.spark.network.netty.NettyBlockTransferService' could not bind on
port 13001. Attempting port 13002.
22/10/04 19:07:49 WARN Utils: Service
'org.apache.spark.network.netty.NettyBlockTransferService' could not bind on
port 13002. Attempting port 13003.
```

[2]:
```
webreq = spark.table('wmf.webrequest').where('webrequest_source=="upload" and␣
↪uri_host=="upload.wikimedia.org" and year==2022 and month==09 and ((day==08␣
↪and hour==17) or (day==13 and hour==02))')
#webreq.show(1,vertical=True,truncate=False)
#webreq.first().toPandas()
```

[3]:
```
# Need to pivot by cp host as hotlinks are very often localized events
breakdown = (webreq
             .groupby(F.window('dt', '1 minute'),
                      'uri_path',
                      'hostname'
                     )
             .agg(F.sum(F.lit(1)).alias('rpm'))
             .withColumn('approx_rps',
                         F.col('rpm')
                         / 60 # secs/minute
                        )
             .where('(approx_rps > 26)')
             .sort(F.desc('approx_rps'))
            )
breakdown.cache()
#breakdown.select('hostname').summary().show()
```

[3]:
```
DataFrame[window: struct<start:timestamp,end:timestamp>, uri_path: string,
hostname: string, rpm: bigint, approx_rps: double]
```

[6]:
```
breakdown.groupby('uri_path').max('approx_rps').sort(F.desc('max(approx_rps)')).
↪toPandas()#show(100,truncate=False)
```

[6]:

	uri_path	max(approx_rps)
0	/wikipedia/commons/c/c2/Two_Bangladeshi_girls_…	999.683333
1	/wikipedia/commons/2/28/Aaj_tak_logo.png	243.933333
2	/wikipedia/commons/thumb/2/26/WLM_Logo_India.s…	56.883333
3	/wikipedia/en/thumb/a/a6/Major_League_Baseball…	46.000000

4	/wikipedia/en/thumb/5/5a/Motor_Trend.svg/300px…	42.166667
5	/wikipedia/commons/thumb/b/b1/CNN.svg/175px-CN…	39.083333
6	/wikipedia/foundation/2/20/CloseWindow19x19.png	38.283333
7	/wikipedia/en/thumb/e/e5/WAAY-TV_2017_logo.png…	35.766667
8	/wikipedia/en/0/08/KATU_logo.png	35.533333
9	/wikipedia/en/d/d1/Image_not_available.png	34.183333
10	/wikipedia/en/thumb/9/9b/MovieMax_logo.svg/192…	33.983333
11	/wikipedia/it/thumb/2/25/RTL_102_5_logo.svg/12…	33.716667
12	/wikipedia/commons/thumb/c/ca/Wiki_Loves_Monum…	33.450000
13	/wikipedia/commons/thumb/b/b2/%D0%A6%D0%B5%D1%…	32.833333
14	/wikipedia/commons/thumb/c/ca/Wiki_Loves_Monum…	32.600000
15	/wikipedia/en/thumb/2/2e/KTVK_logo_2013.png/22…	31.483333
16	/wikipedia/commons/thumb/c/ca/Wiki_Loves_Monum…	29.400000
17	/wikipedia/en/4/48/DuckTales_%28Main_title%29.jpg	28.733333
18	/wikipedia/en/6/69/Sony_Movie_Channel_logo.png	28.600000
19	/wikipedia/commons/thumb/b/b6/Queen_Elizabeth_…	28.516667
20	/wikipedia/commons/thumb/b/b6/Queen_Elizabeth_…	27.233333