```
>>> df.select(split(col("email"),'@')[1].alias("domain")).groupBy("domain").coun
t().withColumn("rank",rank().over(windowSpec)).show(50)
+--------------------+-----+----+

|              domain|count|rank|
+--------------------+-----+----+
|           gmail.com| 8209|   1|
|          apache.org|  451|   2|
|          redhat.com|  278|   3|
|         hotmail.com|  176|   4|
|             163.com|  161|   5|
|          debian.org|  146|   6|
|           yahoo.com|  120|   7|
|              qq.com|  118|   8|
|    googlegroups.com|  105|   9|
|         outlook.com|  104|  10|
|          google.com|  104|  10|
|             gnu.org|  102|  12|
|         freebsd.org|   83|  13|
|              gmx.de|   70|  14|
|      protonmail.com|   69|  15|
|           intel.com|   65|  16|
|          oracle.com|   51|  17|
|      googlemail.com|   44|  18|
|             web.de|   44|  18|
|            gmx.net|   38|  20|
|            126.com|   36|  21|
|        mirantis.com|   35|  22|
|             free.fr|   34|  23|
|           posteo.de|   32|  24|
|         comcast.net|   28|  25|
|       linux.ibm.com|   28|  25|
|          kernel.org|   27|  27|
|         foxmail.com|   27|  27|
|          huawei.com|   27|  27|
|    openjdk.java.net|   27|  27|
|          icloud.com|   26|  31|
|           yandex.ru|   26|  31|
|         t-online.de|   25|  33|
|          python.org|   25|  33|
|          riseup.net|   24|  35|
|             gmx.com|   24|  35|
|             arm.com|   24|  35|
|syzkaller.appspot...|   23|  38|
|          vmware.com|   23|  38|
|           pobox.com|   23|  38|
|       ververica.com|   22|  41|
|       canonical.com|   22|  41|
|    lists.debian.org|   21|  43|
|        ericsson.com|   21|  43|
|            live.com|   21|  43|
|             isc.org|   18|  46|
|         disroot.org|   18|  46|
|     linux.intel.com|   18|  46|
|           nokia.com|   17|  49|
|          linaro.org|   17|  49|
+--------------------+-----+----+
only showing top 50 rows

>>> df.select(reverse(split(col("email"),'\.'))[0].alias("TLD")).groupBy("TLD").
```

```
count().withColumn("rank",rank().over(windowSpec)).show(50)
+----+-----+----+
| TLD|count|rank|
+----+-----+----+
| com|13013|   1|
| org| 2053|   2|
| net|  762|   3|
|  de|  742|   4|
| edu|  301|   5|
|  fr|  217|   6|
|  uk|  206|   7|
|  io|  156|   8|
|  it|  133|   9|
|  ca|  126|  10|
|  ch|  106|  11|
|  ru|   91|  12|
|  nl|   84|  13|
|  au|   81|  14|
|  eu|   75|  15|
|  at|   73|  16|
|  se|   58|  17|
|  pl|   55|  18|
|  jp|   54|  19|
|  cn|   51|  20|
|  us|   50|  21|
|  in|   50|  21|
|  br|   48|  23|
|  cz|   46|  24|
|  ai|   43|  25|
|  nz|   41|  26|
|  be|   36|  27|
|  co|   34|  28|
|  es|   33|  29|
| gov|   31|  30|
|  me|   31|  30|
|  no|   28|  32|
|  dk|   27|  33|
|  fi|   27|  33|
|name|   26|  35|
|  cc|   22|  36|
|info|   21|  37|
| biz|   16|  38|
| dev|   15|  39|
|  ie|   14|  40|
|  pt|   13|  41|
|  hu|   13|  41|
|  gr|   13|  41|
|  ro|   12|  44|
|  tr|   12|  44|
| xyz|   12|  44|
|  fm|   11|  47|
|  sk|   10|  48|
|  mx|   10|  48|
|tech|   10|  48|
+----+-----+----+
only showing top 50 rows

>>> df.count()
19605
```

```
>>> df2.groupBy("word").count().withColumn("rank",rank().over(windowSpec)).show(
100)
+--------------+-----+----+

|          word|count|rank|
+--------------+-----+----+
|       message|31380|   1|
|          send|26530|   2|
|          2022|22459|   3|
|   unsubscribe|21914|   4|
|          file|21247|   5|
|          list|20793|   6|
|       mailing|16356|   7|
|        100644|16002|   8|
|          diff|15964|   9|
|          2021|15625|  10|
|           jan|15418|  11|
|        return|15039|  12|
|         email|14846|  13|
|          mail|14579|  14|
|          code|14077|  15|
|            pm|13948|  16|
|          data|13618|  17|
|           add|13542|  18|
|         group|13416|  19|
|         error|13055|  20|
|       version|12785|  21|
|           set|12745|  22|
|       problem|12590|  23|
|           feb|12536|  24|
|          time|12450|  25|
|        apache|11774|  26|
|      received|11601|  27|
|        struct|11251|  28|
|    discussion|11173|  29|
|          case|11140|  30|
|        google|11130|  31|
|         flink|11120|  32|
|          view|11099|  33|
|         visit|10962|  34|
|         files|10811|  35|
|        groups|10808|  36|
|           web|10763|  37|
|    subscribed|10631|  38|
|     receiving|10595|  39|
|       running|10560|  40|
|         issue|10460|  41|
|       request|10367|  42|
|         check|10363|  43|
|          test|10177|  44|
|          work|10151|  45|
|        emails|10135|  46|
|        change| 9884|  47|
|           int| 9837|  48|
|       support| 9833|  49|
|          read| 9548|  50|
|          user| 9527|  51|
|      specific| 9449|  52|
|           log| 9408|  53|
|          info| 8723|  54|
```

```
|        server| 8627|  55|
|          type| 8598|  56|
|       address| 8516|  57|
|          2020| 8499|  58|
|      function| 8389|  59|
|           job| 8349|  60|
|        memory| 8327|  61|
|        static| 7872|  62|
|       contact| 7759|  63|
|           url| 7711|  64|
|       release| 7651|  65|
|        create| 7640|  66|
|          long| 7595|  67|
|          void| 7438|  68|
|           git| 7336|  69|
|        device| 7153|  70|
|            g1| 7011|  71|
|          good| 7005|  72|
|        github| 6852|  73|
|           dec| 6719|  74|
|       package| 6629|  75|
|         start| 6620|  76|
|           bug| 6598|  77|
|       default| 6452|  78|
|           mar| 6445|  79|
|        source| 6417|  80|
|        python| 6415|  81|
|         build| 6159|  82|
| configuration| 6138|  83|
|          pull| 6137|  84|
|       queries| 6078|  85|
|        update| 6077|  86|
|        failed| 5977|  87|
|           kvm| 5941|  88|
|         guest| 5885|  89|
|       command| 5861|  90|
|infrastructure| 5805|  91|
|      messages| 5776|  92|
|            00| 5775|  93|
|        kernel| 5713|  94|
|        number| 5705|  95|
|     automated| 5701|  96|
|       respond| 5688|  97|
|       current| 5683|  98|
|       provide| 5376|  99|
|    additional| 5357| 100|
+--------------+-----+----+
only showing top 100 rows

>>> df2.count()
4651611
```