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

|              domain|count|rank|
+--------------------+-----+----+
|           gmail.com| 7903|   1|
|          apache.org|  424|   2|
|          redhat.com|  268|   3|
|         hotmail.com|  164|   4|
|             163.com|  150|   5|
|          debian.org|  137|   6|
|           yahoo.com|  117|   7|
|              qq.com|  106|   8|
|    googlegroups.com|  105|   9|
|         outlook.com|  102|  10|
|             gnu.org|  100|  11|
|          google.com|   97|  12|
|         freebsd.org|   83|  13|
|      protonmail.com|   66|  14|
|           intel.com|   65|  15|
|              gmx.de|   62|  16|
|          oracle.com|   47|  17|
|              web.de|   42|  18|
|      googlemail.com|   41|  19|
|             gmx.net|   36|  20|
|        mirantis.com|   35|  21|
|             free.fr|   34|  22|
|             126.com|   32|  23|
|           posteo.de|   29|  24|
|          huawei.com|   27|  25|
|         comcast.net|   27|  25|
|          kernel.org|   27|  25|
|    openjdk.java.net|   27|  25|
|          icloud.com|   26|  29|
|       linux.ibm.com|   26|  29|
|          python.org|   25|  31|
|           yandex.ru|   25|  31|
|             gmx.com|   24|  33|
|          riseup.net|   23|  34|
|         t-online.de|   23|  34|
|       ververica.com|   22|  36|
|syzkaller.appspot...|   22|  36|
|           pobox.com|   22|  36|
|    lists.debian.org|   21|  39|
|       canonical.com|   21|  39|
|             arm.com|   21|  39|
|          vmware.com|   21|  39|
|            live.com|   21|  39|
|         foxmail.com|   20|  44|
|     linux.intel.com|   18|  45|
|           nokia.com|   17|  46|
|             isc.org|   17|  46|
|          linaro.org|   17|  46|
|          amazon.com|   17|  46|
|             cern.ch|   17|  46|
+--------------------+-----+----+
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|12477|   1|
| org| 1983|   2|
| net|  723|   3|
|  de|  702|   4|
| edu|  286|   5|
|  fr|  205|   6|
|  uk|  193|   7|
|  io|  148|   8|
|  it|  127|   9|
|  ca|  117|  10|
|  ch|  100|  11|
|  ru|   87|  12|
|  au|   80|  13|
|  nl|   77|  14|
|  eu|   72|  15|
|  at|   70|  16|
|  se|   55|  17|
|  pl|   53|  18|
|  jp|   51|  19|
|  in|   50|  20|
|  us|   49|  21|
|  cn|   45|  22|
|  br|   45|  22|
|  cz|   45|  22|
|  nz|   38|  25|
|  be|   35|  26|
|  ai|   34|  27|
| gov|   31|  28|
|  co|   31|  28|
|  es|   30|  30|
|  me|   27|  31|
|  fi|   27|  31|
|name|   26|  33|
|  no|   26|  33|
|  dk|   26|  33|
|  cc|   21|  36|
|info|   20|  37|
| dev|   14|  38|
|  ie|   14|  38|
| biz|   14|  38|
|  gr|   13|  41|
|  tr|   12|  42|
| xyz|   12|  42|
|  fm|   11|  44|
|  ro|   11|  44|
|  pt|   11|  44|
|  sk|   10|  47|
|  hu|   10|  47|
|tech|   10|  47|
|  pk|    8|  50|
+----+-----+----+
only showing top 50 rows

>>> df.count()
18755
```

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

|         word|count|rank|
+-------------+-----+----+
|      message|28796|   1|
|         send|25714|   2|
|  unsubscribe|20906|   3|
|         2022|19696|   4|
|         file|19650|   5|
|         list|19502|   6|
|      mailing|15676|   7|
|         2021|15589|   8|
|          jan|15365|   9|
|       100644|14259|  10|
|         mail|14189|  11|
|         diff|14173|  12|
|        email|14148|  13|
|       return|13829|  14|
|         code|13198|  15|
|           pm|13172|  16|
|         data|12785|  17|
|        group|12566|  18|
|          feb|12483|  19|
|      problem|11961|  20|
|      version|11836|  21|
|        error|11821|  22|
|         time|11663|  23|
|          set|11639|  24|
|          add|11609|  25|
|     received|11090|  26|
|        flink|11008|  27|
|   discussion|10692|  28|
|       google|10641|  29|
|         view|10609|  30|
|        visit|10482|  31|
|       groups|10347|  32|
|         case|10326|  33|
|          web|10256|  34|
|       struct|10215|  35|
|   subscribed|10189|  36|
|       apache|10145|  37|
|    receiving|10078|  38|
|      running|10023|  39|
|        files| 9885|  40|
|       emails| 9702|  41|
|         work| 9440|  42|
|         test| 9406|  43|
|        issue| 9343|  44|
|      support| 8998|  45|
|          int| 8993|  46|
|      request| 8986|  47|
|        check| 8981|  48|
|       change| 8915|  49|
|         user| 8748|  50|
|         read| 8661|  51|
|         2020| 8482|  52|
|         info| 8469|  53|
|          job| 8170|  54|
```

```
|     specific| 8162|  55|
|      address| 8120|  56|
|       server| 8104|  57|
|         type| 8098|  58|
|          log| 7864|  59|
|     function| 7748|  60|
|       memory| 7482|  61|
|       static| 7205|  62|
|         long| 7110|  63|
|       create| 7019|  64|
|         void| 6869|  65|
|      release| 6757|  66|
|          dec| 6706|  67|
|         good| 6592|  68|
|       device| 6578|  69|
|      contact| 6503|  70|
|          url| 6453|  71|
|       python| 6235|  72|
|        start| 6231|  73|
|      package| 6154|  74|
|          git| 6119|  75|
|      default| 6044|  76|
|       source| 5948|  77|
|          bug| 5698|  78|
|       github| 5662|  79|
|           00| 5638|  80|
|        build| 5615|  81|
|       failed| 5602|  82|
|     messages| 5554|  83|
|      command| 5523|  84|
|        guest| 5393|  85|
|          kvm| 5356|  86|
|       number| 5331|  87|
|       update| 5325|  88|
|configuration| 5239|  89|
|         pull| 5066|  90|
|       kernel| 5036|  91|
|      current| 5022|  92|
|      provide| 5008|  93|
|      queries| 5005|  94|
|          bit| 4930|  95|
|       cygwin| 4916|  96|
|        table| 4884|  97|
|        write| 4852|  98|
|      working| 4815|  99|
|   additional| 4749| 100|
+-------------+-----+----+
only showing top 100 rows

>>> df2.count()
4284297
```