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

|              domain|count|rank|
+--------------------+-----+----+
|           gmail.com| 9523|   1|
|          apache.org|  557|   2|
|          redhat.com|  314|   3|
|         hotmail.com|  196|   4|
|             163.com|  188|   5|
|          debian.org|  175|   6|
|           yahoo.com|  139|   7|
|              qq.com|  137|   8|
|         outlook.com|  121|   9|
|          google.com|  121|   9|
|             gnu.org|  110|  11|
|    googlegroups.com|  109|  12|
|              gmx.de|   88|  13|
|         freebsd.org|   86|  14|
|      protonmail.com|   81|  15|
|           intel.com|   68|  16|
|          oracle.com|   63|  17|
|              web.de|   56|  18|
|      googlemail.com|   51|  19|
|             gmx.net|   47|  20|
|             126.com|   46|  21|
|             free.fr|   42|  22|
|        mirantis.com|   36|  23|
|           posteo.de|   34|  24|
|         foxmail.com|   34|  24|
|          huawei.com|   33|  26|
|          icloud.com|   33|  26|
|         comcast.net|   32|  28|
|        cloudera.com|   30|  29|
|           yandex.ru|   29|  30|
|          riseup.net|   28|  31|
|          kernel.org|   28|  31|
|    openjdk.java.net|   28|  31|
|         t-online.de|   28|  31|
|          python.org|   28|  31|
|       linux.ibm.com|   28|  31|
|              gmx.com|   27|  37|
|syzkaller.appspot...|   27|  37|
|              arm.com|   26|  39|
|          disroot.org|   26|  39|
|       canonical.com|   25|  41|
|           vmware.com|   25|  41|
|            pobox.com|   24|  43|
|             live.com|   23|  44|
|       ververica.com|   22|  45|
|             isc.org|   22|  45|
|        ericsson.com|   22|  45|
|    lists.debian.org|   21|  48|
|           amazon.com|   19|  49|
|     linux.intel.com|   19|  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|15048|   1|
| org| 2322|   2|
| net|  869|   3|
|  de|  856|   4|
| edu|  352|   5|
|  fr|  247|   6|
|  uk|  242|   7|
|  io|  179|   8|
|  it|  157|   9|
|  ca|  150|  10|
|  ch|  117|  11|
|  ru|  105|  12|
|  nl|  102|  13|
|  au|   93|  14|
|  eu|   82|  15|
|  at|   81|  16|
|  se|   70|  17|
|  pl|   63|  18|
|  cn|   62|  19|
|  jp|   61|  20|
|  br|   60|  21|
|  in|   58|  22|
|  cz|   57|  23|
|  us|   53|  24|
|  ai|   52|  25|
|  nz|   47|  26|
|  be|   44|  27|
|  me|   42|  28|
| gov|   38|  29|
|  co|   37|  30|
|  es|   37|  30|
|  no|   34|  32|
|  fi|   33|  33|
|name|   31|  34|
|  dk|   31|  34|
|  cc|   25|  36|
|info|   25|  36|
| dev|   19|  38|
|  hu|   18|  39|
| xyz|   17|  40|
|  gr|   17|  40|
| biz|   17|  40|
|  ie|   16|  43|
|  pt|   15|  44|
|  tr|   14|  45|
|  fm|   13|  46|
|  ar|   13|  46|
|  ro|   12|  48|
|  sk|   11|  49|
|  sg|   11|  49|
+----+-----+----+
only showing top 50 rows

>>> df.count()
22629
```

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

|          word|count|rank|
+--------------+-----+----+
|       message|40467|   1|
|          2022|31370|   2|
|          send|29323|   3|
|          file|26890|   4|
|   unsubscribe|25108|   5|
|          list|24695|   6|
|     receiving|24221|   7|
|          diff|20917|   8|
|        100644|20671|   9|
|        return|18609|  10|
|       mailing|18421|  11|
|          code|17753|  12|
|        apache|17749|  13|
|         email|17643|  14|
|           add|17036|  15|
|          data|16616|  16|
|       version|16528|  17|
|           set|16388|  18|
|            pm|16386|  19|
|         error|16349|  20|
|          mail|16220|  21|
|          2021|15723|  22|
|         group|15688|  23|
|           jan|15597|  24|
|          time|15056|  25|
|       problem|14599|  26|
|           log|14492|  27|
|      specific|14469|  28|
|        struct|14468|  29|
|         issue|14105|  30|
|         check|14060|  31|
|         files|13821|  32|
|          case|13475|  33|
|      received|13413|  34|
|       request|13395|  35|
|          test|13162|  36|
|    discussion|12955|  37|
|          view|12789|  38|
|         visit|12786|  39|
|        google|12759|  40|
|       running|12689|  41|
|       support|12598|  42|
|           feb|12595|  43|
|        change|12578|  44|
|           url|12544|  45|
|          work|12499|  46|
|           web|12464|  47|
|           git|12385|  48|
|       contact|12374|  49|
|        groups|12361|  50|
|           mib|12357|  51|
|           int|12264|  52|
|    subscribed|12131|  53|
|          read|12067|  54|
```

```
|        emails|11725|  55|
|        github|11665|  56|
|         flink|11601|  57|
|          user|11513|  58|
|        server|10565|  59|
|        memory|10485|  60|
|       queries|10435|  61|
|          type|10295|  62|
|        static|10290|  63|
|      function|10276|  64|
|     automated|10170|  65|
|       release|10165|  66|
|infrastructure|10111|  67|
|       respond| 9956|  68|
|       address| 9878|  69|
|          info| 9778|  70|
|          void| 9540|  71|
|        create| 9520|  72|
|          long| 9420|  73|
|           mar| 9254|  74|
|           job| 9098|  75|
|           bug| 9018|  76|
|        device| 8966|  77|
|          pull| 8812|  78|
|          good| 8637|  79|
|       default| 8570|  80|
|            pr| 8552|  81|
|          2020| 8537|  82|
|       package| 8073|  83|
|         build| 8051|  84|
|           apr| 8042|  85|
|         start| 8041|  86|
|       current| 8000|  87|
|         class| 7840|  88|
|        update| 7827|  89|
|        failed| 7818|  90|
|    additional| 7816|  91|
|           kvm| 7726|  92|
|        source| 7630|  93|
|         guest| 7454|  94|
|        kernel| 7390|  95|
|            00| 7353|  96|
| configuration| 7300|  97|
|          jira| 7196|  98|
|        number| 7145|  99|
|        python| 7140| 100|
+--------------+-----+----+
only showing top 100 rows

>>> df2.count()
5858127
```