```
>>> df.select(split(col("email"),'@')[1].alias("domain"),"count").groupBy("domai
n").count().orderBy("count",ascending=False).show(50)
+--------------------+-----+

|              domain|count|
+--------------------+-----+
|           gmail.com| 4843|
|          redhat.com|  161|
|          apache.org|  132|
|             163.com|  112|
|         hotmail.com|   87|
|             gnu.org|   86|
|         freebsd.org|   76|
|              qq.com|   65|
|          google.com|   61|
|           yahoo.com|   59|
|         outlook.com|   58|
|           intel.com|   55|
|          debian.org|   42|
|        mirantis.com|   35|
|          oracle.com|   30|
|    openjdk.java.net|   27|
|              gmx.de|   27|
|    googlegroups.com|   25|
|          python.org|   24|
|             126.com|   23|
|             gmx.net|   21|
|          huawei.com|   21|
|           yandex.ru|   21|
|       ververica.com|   20|
|    lists.debian.org|   20|
|       linux.ibm.com|   19|
|              web.de|   19|
|             arm.com|   18|
|          golang.org|   16|
|      googlemail.com|   16|
|     linux.intel.com|   16|
|         t-online.de|   16|
|          icloud.com|   15|
|           pobox.com|   15|
|syzkaller.appspot...|   15|
|      protonmail.com|   15|
|          vmware.com|   14|
|           cisco.com|   14|
|             free.fr|   14|
|       openstack.org|   14|
|           nokia.com|   13|
|             gmx.com|   13|
|          kernel.org|   13|
|          linaro.org|   13|
|             hpe.com|   13|
|          cygwin.com|   13|
|         comcast.net|   12|
|       canonical.com|   12|
|        ericsson.com|   12|
|           posteo.de|   11|
+--------------------+-----+
only showing top 50 rows

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

```
("TLD").count().orderBy("count",ascending=False).show(50)
+----+-----+
| TLD|count|
+----+-----+
| com| 7580|
| org|  988|
| net|  391|
|  de|  343|
| edu|  118|
|  uk|   99|
|  io|   89|
|  fr|   77|
|  ca|   58|
|  ru|   57|
|  au|   54|
|  ch|   49|
|  it|   48|
|  nl|   44|
|  eu|   41|
|  at|   40|
|  jp|   38|
|  se|   34|
|  in|   33|
|  cn|   30|
|  us|   29|
|  cz|   28|
|  pl|   26|
|  br|   26|
|  nz|   18|
|  ai|   18|
|  be|   18|
|name|   17|
|  me|   17|
|  co|   16|
|  es|   15|
|  no|   15|
|  fi|   12|
|info|   11|
|  dk|   11|
|  cc|   10|
| gov|   10|
|  gr|    9|
|  hu|    8|
| dev|    8|
| xyz|    8|
|  sk|    7|
|  kr|    7|
|  ie|    7|
| biz|    7|
|  tr|    6|
|  fm|    6|
|  mx|    6|
|  ro|    5|
|  pt|    5|
+----+-----+
only showing top 50 rows

>>> df.count()
10841
```

```
>>> df2.select("*").groupBy("word").count().orderBy("count",ascending=False).sho
w(100)
+-------------+-----+

|         word|count|
+-------------+-----+
|         send|15531|
|         2021|14948|
|  unsubscribe|14271|
|      message|12244|
|         list|11135|
|      mailing|10338|
|            0|10129|
|        flink| 9891|
|         file| 9525|
|            1| 9487|
|         mail| 9342|
|        email| 8932|
|         2020| 8393|
|           pm| 8389|
|      problem| 8226|
|        group| 8040|
|         data| 7987|
|         code| 7794|
|     received| 7612|
|   discussion| 7354|
|       google| 7241|
|       groups| 7105|
|         view| 7004|
|        visit| 6973|
|          web| 6966|
|       return| 6960|
|   subscribed| 6954|
|            2| 6940|
|    receiving| 6859|
|         time| 6833|
|          job| 6805|
|       emails| 6746|
|          dec| 6314|
|        error| 6087|
|          set| 6000|
|      version| 5887|
|       100644| 5815|
|      running| 5744|
|         diff| 5576|
|         type| 5389|
|         info| 5359|
|         work| 5269|
|         case| 5125|
|       python| 4988|
|        files| 4898|
|          add| 4790|
|       struct| 4609|
|          nov| 4527|
|         user| 4489|
|       cygwin| 4368|
|         read| 4345|
|          int| 4201|
|      support| 4061|
|            3| 4056|
```

```
|       memory| 4038|
|       change| 3994|
|     function| 3989|
|        issue| 3903|
|       server| 3869|
|       create| 3773|
|         good| 3587|
|      address| 3552|
|       source| 3541|
|         test| 3507|
|        check| 3444|
|        start| 3366|
|         long| 3282|
|            4| 3199|
|      default| 3189|
|       static| 3124|
|   connection| 3115|
|configuration| 3081|
|         void| 3074|
|      freebsd| 3038|
|      release| 3006|
|       number| 2983|
|        class| 2975|
|      package| 2941|
|        write| 2896|
|       device| 2850|
|      provide| 2819|
|        table| 2802|
|       client| 2787|
|       output| 2772|
|      command| 2762|
|       import| 2730|
|      cluster| 2729|
|         task| 2701|
|        works| 2630|
|            5| 2629|
|      current| 2625|
|      process| 2624|
|      windows| 2607|
|         post| 2605|
|          bit| 2603|
|        space| 2507|
|          jun| 2491|
|  application| 2480|
|        build| 2469|
|        based| 2462|
+-------------+-----+
only showing top 100 rows

>>> df2.count()
2224313
```