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

|              domain|count|
+--------------------+-----+
|           gmail.com| 5076|
|          redhat.com|  172|
|          apache.org|  166|
|             163.com|  114|
|         hotmail.com|   92|
|             gnu.org|   88|
|    googlegroups.com|   78|
|         freebsd.org|   77|
|              qq.com|   68|
|          google.com|   62|
|         outlook.com|   61|
|           yahoo.com|   61|
|           intel.com|   56|
|          debian.org|   54|
|        mirantis.com|   35|
|          oracle.com|   33|
|              gmx.de|   31|
|    openjdk.java.net|   27|
|             126.com|   24|
|          python.org|   24|
|              web.de|   21|
|      googlemail.com|   21|
|       ververica.com|   21|
|             gmx.net|   21|
|           yandex.ru|   21|
|    lists.debian.org|   21|
|          huawei.com|   21|
|       linux.ibm.com|   20|
|             arm.com|   18|
|     linux.intel.com|   16|
|      protonmail.com|   16|
|          golang.org|   16|
|         t-online.de|   16|
|             free.fr|   16|
|          icloud.com|   16|
|syzkaller.appspot...|   15|
|           pobox.com|   15|
|       canonical.com|   15|
|             gmx.com|   15|
|       openstack.org|   14|
|          vmware.com|   14|
|          kernel.org|   14|
|           cisco.com|   14|
|           posteo.de|   13|
|           nokia.com|   13|
|         comcast.net|   13|
|          linaro.org|   13|
|             hpe.com|   13|
|          cygwin.com|   13|
|        ericsson.com|   13|
+--------------------+-----+
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| 8013|
| org| 1289|
| net|  418|
|  de|  377|
| edu|  133|
|  uk|  109|
|  io|   92|
|  fr|   90|
|  ca|   62|
|  ru|   59|
|  au|   57|
|  it|   54|
|  ch|   51|
|  eu|   44|
|  nl|   44|
|  at|   43|
|  jp|   39|
|  in|   37|
|  se|   34|
|  cn|   32|
|  us|   31|
|  pl|   31|
|  cz|   31|
|  br|   27|
|  ai|   21|
|  nz|   19|
|  be|   19|
|  co|   19|
|  es|   18|
|name|   18|
|  me|   18|
|  no|   15|
|  fi|   13|
|  cc|   12|
|  dk|   12|
|info|   12|
| gov|   11|
| dev|    9|
| xyz|    9|
|  gr|    9|
|  ie|    8|
|  sk|    8|
|  hu|    8|
|  kr|    7|
|  fm|    7|
|  mx|    7|
|  tr|    7|
| biz|    7|
|  pt|    6|
|  is|    6|
+----+-----+
only showing top 50 rows

>>> df.count()
11762
```

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

|         word|count|
+-------------+-----+
|         send|20210|
|      message|16121|
|         2021|15124|
|  unsubscribe|14922|
|         list|12340|
|      mailing|12028|
|         mail|11369|
|         file|10184|
|        flink| 9981|
|        email| 9358|
|           pm| 8777|
|      problem| 8485|
|        group| 8420|
|         2020| 8396|
|         data| 8244|
|         code| 8071|
|     received| 7940|
|       google| 7602|
|   discussion| 7587|
|       groups| 7411|
|   subscribed| 7394|
|        visit| 7384|
|         view| 7247|
|       return| 7242|
|          web| 7181|
|         time| 7094|
|    receiving| 7062|
|       emails| 6926|
|          job| 6894|
|        error| 6550|
|          jan| 6459|
|          dec| 6455|
|          set| 6359|
|      version| 6224|
|       100644| 6038|
|      address| 5940|
|      running| 5931|
|         info| 5910|
|         diff| 5815|
|         case| 5807|
|         type| 5606|
|          add| 5585|
|         work| 5559|
|        files| 5131|
|       python| 5049|
|         2022| 4964|
|       server| 4959|
|         user| 4902|
|       struct| 4834|
|         read| 4570|
|          nov| 4533|
|       change| 4469|
|       cygwin| 4421|
|     messages| 4403|
|          int| 4392|
|      support| 4289|
|       memory| 4183|
|        issue| 4148|
|     function| 4093|
|        start| 4086|
|       create| 3971|
|      command| 3806|
|         good| 3744|
|        check| 3738|
|         test| 3717|
|         long| 3675|
|       source| 3653|
|      default| 3388|
|       static| 3247|
|         void| 3173|
|      release| 3159|
|configuration| 3159|
|   connection| 3150|
|       number| 3110|
|      package| 3094|
|      freebsd| 3050|
|       device| 3012|
|        class| 3012|
|        write| 2994|
|      provide| 2964|
|       client| 2894|
|      request| 2890|
|       apache| 2876|
|        table| 2872|
|      working| 2847|
|       output| 2846|
|       import| 2813|
|         task| 2806|
|      cluster| 2782|
|          lot| 2756|
|      current| 2747|
|        works| 2746|
|      process| 2739|
|         port| 2723|
|          bit| 2707|
|      windows| 2697|
|       public| 2680|
|           00| 2672|
|         post| 2656|
|        build| 2648|
+-------------+-----+
only showing top 100 rows

>>> df2.count()
2356361
```