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

|              domain|count|
+--------------------+-----+
|           gmail.com| 6689|
|          apache.org|  333|
|          redhat.com|  230|
|             163.com|  137|
|         hotmail.com|  136|
|          debian.org|  103|
|    googlegroups.com|  101|
|             gnu.org|   94|
|           yahoo.com|   94|
|          google.com|   84|
|         outlook.com|   83|
|              qq.com|   81|
|         freebsd.org|   81|
|           intel.com|   60|
|              gmx.de|   57|
|          oracle.com|   41|
|      protonmail.com|   39|
|        mirantis.com|   35|
|              web.de|   34|
|      googlemail.com|   31|
|             gmx.net|   31|
|             126.com|   28|
|    openjdk.java.net|   27|
|             free.fr|   26|
|          python.org|   25|
|       linux.ibm.com|   24|
|           posteo.de|   24|
|          huawei.com|   23|
|           yandex.ru|   23|
|          icloud.com|   22|
|       ververica.com|   21|
|         t-online.de|   21|
|         comcast.net|   21|
|    lists.debian.org|   21|
|          kernel.org|   20|
|       canonical.com|   20|
|           pobox.com|   20|
|             gmx.com|   20|
|         foxmail.com|   19|
|             arm.com|   19|
|            live.com|   18|
|          riseup.net|   18|
|syzkaller.appspot...|   17|
|        ericsson.com|   17|
|     linux.intel.com|   17|
|          golang.org|   16|
|           nokia.com|   16|
|          vmware.com|   15|
|       openstack.org|   15|
|           cisco.com|   14|
+--------------------+-----+
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|10529|
| org| 1704|
| net|  596|
|  de|  583|
| edu|  223|
|  uk|  156|
|  fr|  146|
|  io|  120|
|  it|   90|
|  ca|   90|
|  ch|   78|
|  ru|   72|
|  au|   67|
|  nl|   63|
|  eu|   61|
|  at|   59|
|  se|   47|
|  jp|   46|
|  pl|   44|
|  in|   44|
|  us|   42|
|  cn|   42|
|  cz|   36|
|  br|   35|
|  nz|   29|
|  ai|   27|
|  be|   27|
|  co|   25|
|name|   24|
|  me|   24|
|  es|   23|
|  fi|   22|
|  no|   21|
|  dk|   20|
| gov|   19|
|info|   16|
|  cc|   14|
| dev|   12|
|  gr|   12|
| biz|   11|
|  ie|   11|
|  tr|   11|
| xyz|   11|
|  sk|   10|
|  fm|   10|
|  hu|    9|
|  pt|    9|
|  ro|    8|
|  mx|    8|
|  pk|    7|
+----+-----+
only showing top 50 rows

>>> df.count()
15721
```

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

|         word|count|
+-------------+-----+
|         send|23074|
|      message|21660|
|  unsubscribe|17893|
|         list|15736|
|         2021|15411|
|          jan|15082|
|         file|14571|
|      mailing|13921|
|         mail|12722|
|         2022|11940|
|        email|11754|
|           pm|10854|
|         code|10595|
|        flink|10482|
|        group|10348|
|         data|10303|
|      problem|10222|
|       return| 9537|
|     received| 9456|
|         time| 9253|
|      version| 9201|
|   discussion| 9121|
|       google| 9086|
|        error| 9070|
|         view| 8971|
|        visit| 8894|
|       groups| 8858|
|   subscribed| 8794|
|          set| 8784|
|          web| 8758|
|    receiving| 8541|
|       100644| 8527|
|         2020| 8425|
|         diff| 8391|
|          add| 8343|
|      running| 8296|
|       emails| 8274|
|         case| 8013|
|          job| 7559|
|         info| 7469|
|         test| 7464|
|         work| 7449|
|        files| 7203|
|      address| 6884|
|         user| 6668|
|          dec| 6666|
|         type| 6591|
|        issue| 6578|
|         read| 6550|
|       change| 6469|
|      support| 6432|
|       struct| 6375|
|       server| 6365|
|        check| 6356|
```

```
|       apache| 6248|
|     function| 6065|
|          int| 5916|
|       python| 5628|
|      request| 5595|
|       memory| 5586|
|       create| 5298|
|         good| 5220|
|        start| 5144|
|           00| 5129|
|     messages| 4985|
|      release| 4980|
|         long| 4976|
|     specific| 4883|
|       source| 4751|
|       cygwin| 4739|
|      default| 4686|
|      package| 4674|
|          log| 4659|
|      command| 4622|
|          nov| 4601|
|       failed| 4480|
|       static| 4434|
|         void| 4344|
|       device| 4295|
|        build| 4225|
|configuration| 4220|
|       number| 4216|
|           16| 4208|
|      provide| 3899|
|       output| 3850|
|        write| 3817|
|      working| 3810|
|          bug| 3809|
|      current| 3793|
|          bit| 3746|
|       update| 3693|
|   connection| 3654|
|       client| 3633|
|      contact| 3628|
|        table| 3627|
|        guest| 3599|
|        class| 3558|
|        works| 3550|
|      process| 3540|
|         port| 3503|
+-------------+-----+
only showing top 100 rows

>>> df2.count()
3267497
```