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

|              domain|count|
+--------------------+-----+
|           gmail.com| 4646|
|          redhat.com|  158|
|          apache.org|  129|
|             163.com|  110|
|         hotmail.com|   83|
|             gnu.org|   77|
|         freebsd.org|   76|
|              qq.com|   65|
|          google.com|   59|
|           intel.com|   54|
|           yahoo.com|   54|
|         outlook.com|   50|
|        mirantis.com|   35|
|          oracle.com|   28|
|    openjdk.java.net|   27|
|          python.org|   24|
|    googlegroups.com|   23|
|             126.com|   23|
|          huawei.com|   21|
|           yandex.ru|   21|
|       ververica.com|   20|
|    lists.debian.org|   20|
|              gmx.de|   19|
|       linux.ibm.com|   19|
|             gmx.net|   19|
|             arm.com|   18|
|          debian.org|   16|
|          golang.org|   16|
|     linux.intel.com|   16|
|      googlemail.com|   15|
|              web.de|   15|
|          icloud.com|   14|
|          vmware.com|   14|
|syzkaller.appspot...|   14|
|           pobox.com|   14|
|       openstack.org|   14|
|      protonmail.com|   14|
|           cisco.com|   14|
|          linaro.org|   13|
|          cygwin.com|   13|
|             hpe.com|   13|
|          kernel.org|   13|
|         t-online.de|   13|
|           nokia.com|   12|
|       canonical.com|   12|
|        ericsson.com|   12|
|             gmx.com|   11|
|         foxmail.com|   10|
|       rackspace.com|   10|
|         comcast.net|   10|
+--------------------+-----+
only showing top 50 rows
```

```
# domains TLD statistics
>>> df.select(reverse(split(col("email"),'\.'))[0].alias("TLD"),"count").groupBy
("TLD").count().orderBy("count",ascending=False).show(50)
+----+-----+
| TLD|count|
+----+-----+
| com| 7266|
| org|  886|
| net|  360|
|  de|  293|
| edu|  109|
|  uk|   92|
|  io|   87|
|  fr|   64|
|  ru|   55|
|  ca|   54|
|  au|   46|
|  nl|   40|
|  ch|   40|
|  it|   38|
|  jp|   36|
|  at|   35|
|  eu|   34|
|  in|   33|
|  cn|   30|
|  se|   30|
|  us|   27|
|  br|   26|
|  pl|   23|
|  cz|   22|
|  ai|   18|
|  nz|   16|
|  be|   16|
|  co|   16|
|  es|   14|
|name|   14|
|  me|   13|
|  fi|   12|
|  cc|   10|
| gov|   10|
|info|   10|
|  no|   10|
|  dk|   10|
|  gr|    9|
|  hu|    8|
|  ie|    7|
|  sk|    7|
| biz|    7|
| xyz|    7|
|  kr|    7|
|  tr|    6|
| dev|    6|
|  mx|    5|
|  is|    5|
|  ro|    5|
|  fm|    5|
+----+-----+
only showing top 50 rows

# total email addresses
```

```
>>> df.count()
10200
```