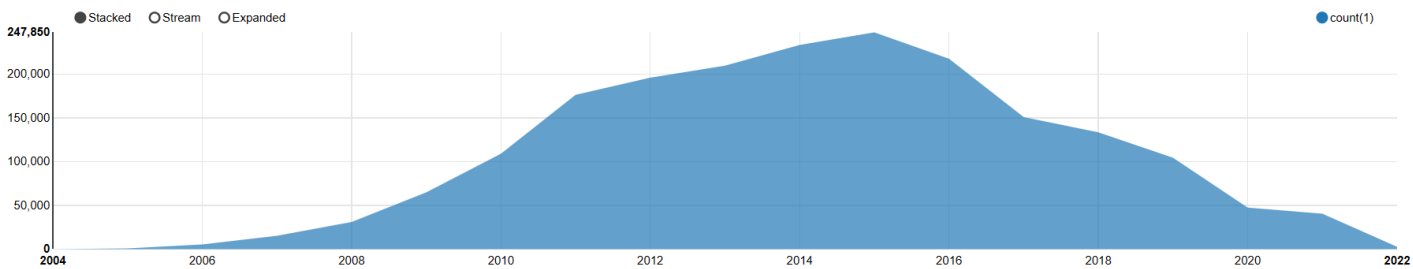
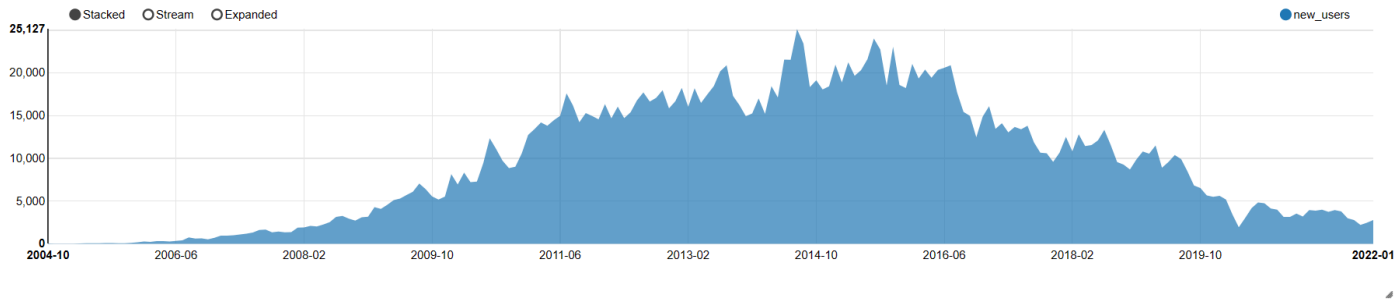
用户统计及分析模块Spark

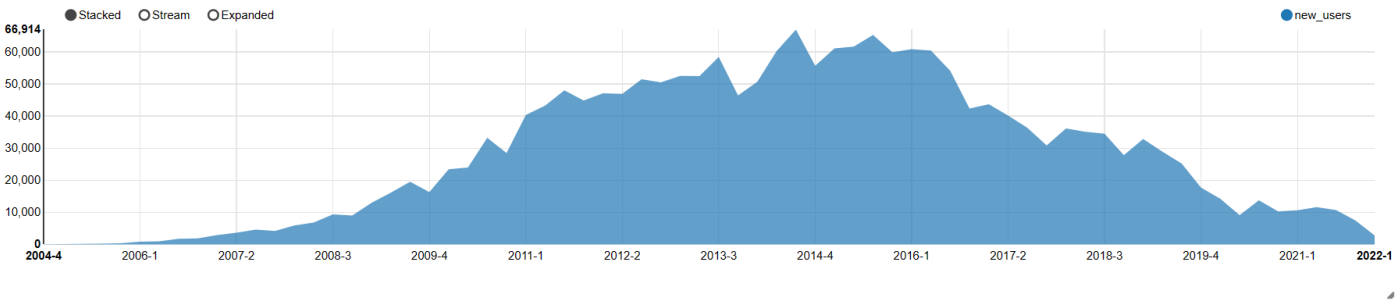
```pyspark
%pyspark

# 编写者:凌晨
# 功能:分析每年加入的用户数量
# 时间:2024.1.3
from pyspark.sql import HiveContext

hc = HiveContext(sc)
result = hc.sql("SELECT year(to_date(user_yelping_since)) as year, COUNT(*) FROM users GROUP BY year(to_date(user_yelping_since)) ORDER BY year DESC")
z.show(result)
```




```pyspark
%pyspark

# 编写者:凌晨
# 功能:分析每月加入的用户数量
# 时间:2024.1.3
from pyspark.sql import HiveContext

hc = HiveContext(sc)
result = hc.sql("SELECT DATE_FORMAT(user_yelping_since, 'yyyy-MM') AS month, COUNT(*) AS new_users FROM users GROUP BY DATE_FORMAT(user_yelping_since, 'yyyy-MM') ORDER BY month")
z.show(result)
```

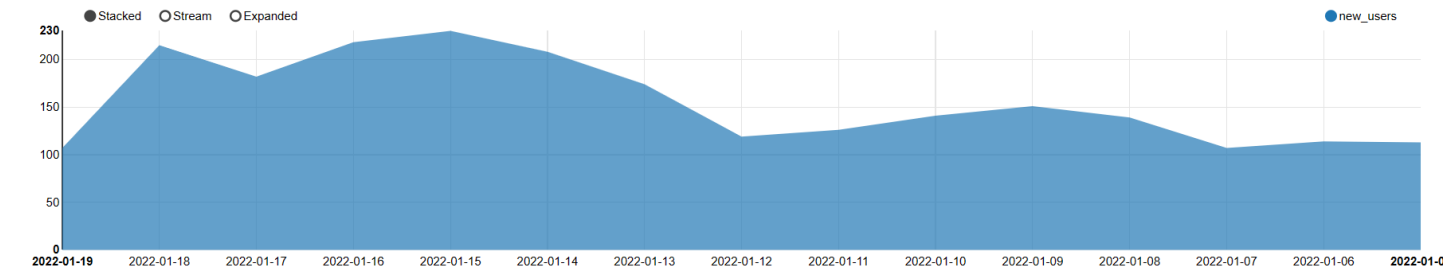



```pyspark

# 编写者:凌晨
# 功能:分析每季度加入的用户数量，每三个月视为一个季度。如2004-4代表2004的第四个季度
# 时间:2024.1.3
from pyspark.sql import HiveContext

hc = HiveContext(sc)
result = hc.sql('''
SELECT
  CONCAT(YEAR(user_yelping_since), '-', CEIL(MONTH(user_yelping_since) / 3)) AS quarter,
  COUNT(*) AS new_users
FROM users
GROUP BY CONCAT(YEAR(user_yelping_since), '-', CEIL(MONTH(user_yelping_since) / 3))
ORDER BY quarter
''')
z.show(result)
```



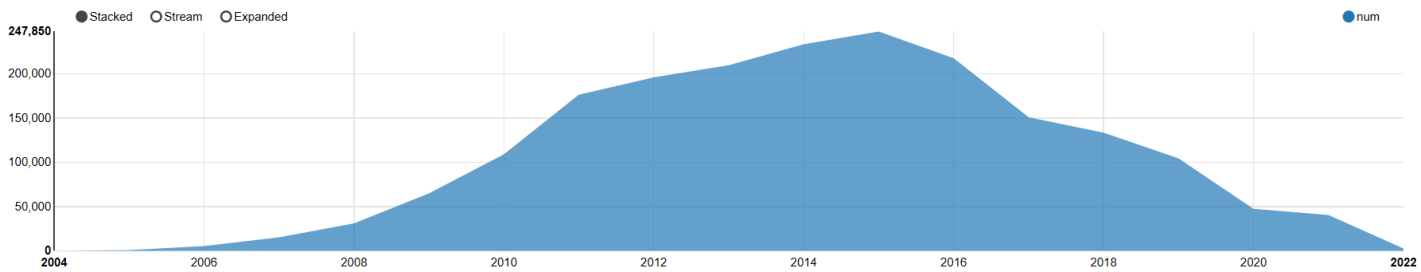

```%pyspark

# 编写者:凌晨
# 功能:分析最近15天内的加入的用户数量
# 时间:2024.1.3
```

```
6
7  from pyspark.sql import HiveContext
8
9  hc = HiveContext(sc)
10 result = hc.sql("SELECT DATE(user_yelping_since) AS date, COUNT(*) AS
   new_users FROM users GROUP BY DATE(user_yelping_since) ORDER BY date DESC
   LIMIT 15")
11 z.show(result)
```



Took 4 sec. Last updated by anonymous at January 03 2024, 3:27:04 PM. (outdated)

```
1  %pyspark
2
3  # 编写者:凌晨
4  # 功能:分析每年加入的用户数量，按照用户数排序
5  # 时间:2024.1.3
6  from pyspark.sql import HiveContext
7
8  hc = HiveContext(sc)
9  result = hc.sql("SELECT year(to_date(user_yelping_since)) as year, COUNT(*) AS
   num FROM users GROUP BY year ORDER BY num DESC")
10 z.show(result)
```




```
1  %pyspark
2
3  # 编写者:凌晨
```

```
# 功能:找出最常见的前20的用户名
# 时间:2024.1.3

from pyspark.sql import HiveContext

hc = HiveContext(sc)
result = hc.sql("SELECT user_name, COUNT(user_name) as name_count FROM users GROUP BY user_name ORDER BY name_count DESC LIMIT 20")

z.show(result)
```

user_name	name_count
John	18719
Michael	16645
David	15967
Chris	14591
Mike	13390
Jennifer	13275
Sarah	11893
Jessica	11688

```
%pyspark

# 编写者:凌晨
# 功能:统计已注册的全部用户数量
# 时间:2024.1.3

from pyspark.sql import HiveContext

hc = HiveContext(sc)
result = hc.sql("SELECT COUNT(*) as num_of_user FROM users")

z.show(result)
```

num_of_user
1987897

```pyspark
%pyspark

# 编写者:钟鹏
# 功能:通过id查找name
# 时间:2024.1.4

from pyspark.sql import HiveContext
from pyspark.sql.functions import *

hcx = HiveContext(sc)

df = hcx.table('users')

user_id = 'Js4nQlGRjS1Bd5vsAQQfgA'
# 需要查找的 user_id

result = df.filter(col('user_id') == user_id).select('user_name')

z.show(result)
```

user_name
Tom