Máster en Data Science. URJC Sede Madrid

Trabajo final - Elasticsearch & Kibana

Javier Llorente Mañas
Carlos Sánchez Vega

2017

Índice

0.1 ¿Qué es "OPNFV"?

OPNFV es un proyecto de código libre y colaborativo que tiene como objetivo crear un marco de referencia, para los proyectos de código libre, para virtualizar servicios en red. Aboga por aprovechar los desarrollos existentes y desarrollar únicamente las partes necesarias para nuevas funcionalidades. De esta manera, se eliminan forks innecesarios de proyectos. OPNFV integra los nuevos desarrollos, despliega, los testea, integra y, finalmente, publica los resultados de forma automatizada e iterativa. La filosofía "DevOps CI/CD" (Continuous Integration and Continuous Deployment) constituye el núcleo de OPNFV. Esencialmente, OPNFV proporciona la estructura para construir comunicaciones virtualizadas, así como el software para gestionarlo.

0.2 Secciones de la memoria

- Creación del índice e importación de los datos
- Visualizaciones basadas en el conjunto de datos
- Conclusiones

0.3 Creación del índice e importación de los datos

1. Se crea un nuevo índice:

```
curl -XPUT -k localhost:9200/git_openfv
```

2. Se crea el mapping:

```
elasticdump
    --input=/home/csanchez/Downloads/opnfv2018/git_opnfv_mapping.json
    --output=http://localhost:9200
--type=mapping --output-index=git_opnfv
    --headers='{"Content-Type":"application/json"}'
```

3. Se importan los datos

```
elasticdump
    --input=/home/csanchez/Downloads/opnfv2018/git_opnfv_data.json
    --output=http://localhost:9200
--type=data --output-index=git_opnfv
    --headers='{"Content-Type":"application/json"}'
```

0.4 Visualizaciones basadas en el conjunto de datos

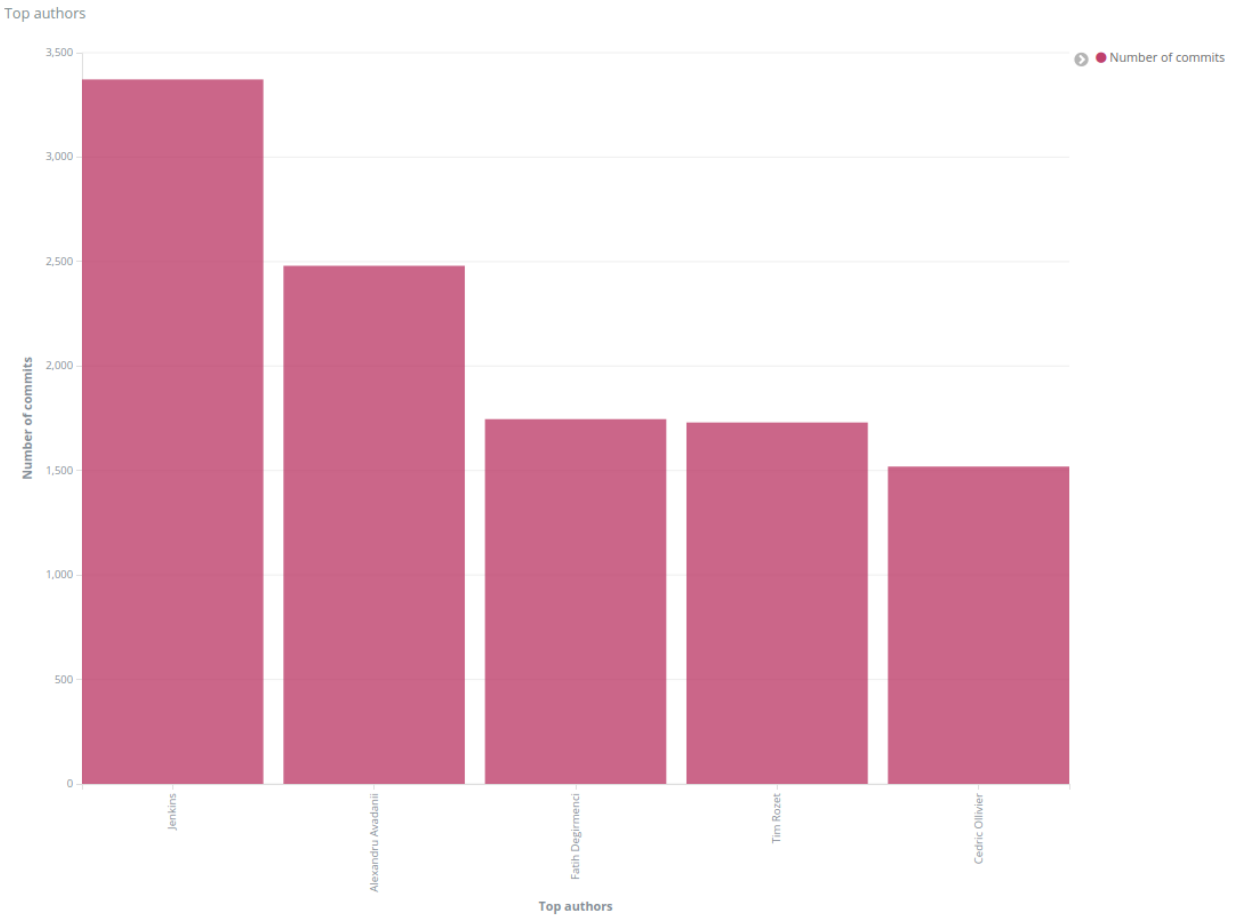
El conjunto de datos corresponde a datos de commits en github de varias compañías de comunicaciones (ZTE, Ericsson…). En el conjunto consta de 4933 commits, tal y como mostramos en el siguiente gráfico:

En el conjunto de datos, se pueden distinguir los siguientes actores:

1. Author: es la persona que creó, inicialmente, el código.
2. Commiter: es la persona que hizo commit en el código inicialmente creado por un autor.

Por tanto, trataremos ambos casos por separado.

0.4.1 Authors



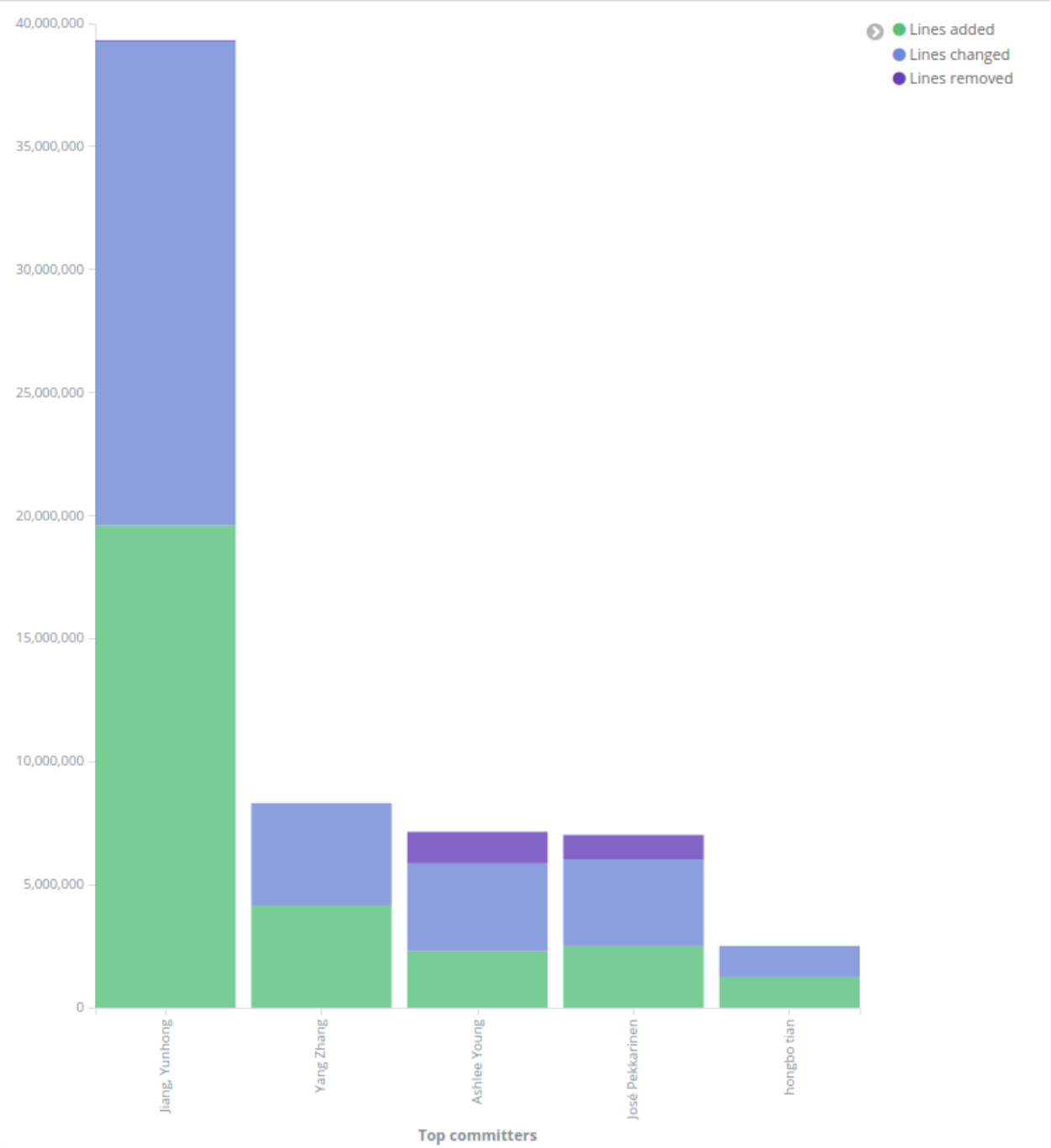
Jenkins es considerada como el mayor autor en el conjunto de datos. Podría ocurrir que "Jenkins" fuera, en realidad, la herramienta de integración continua que usan muchas compañías. Sin enbargo, si buscamos en el campo "Author_bot", podemos ver que todos los campos no son creados automáticamente.

Con lo cual, podríamos pensar que "Jenkins" pudiera corresponderse con el nombre de una persona.

0.4.2 Author organizations and authors

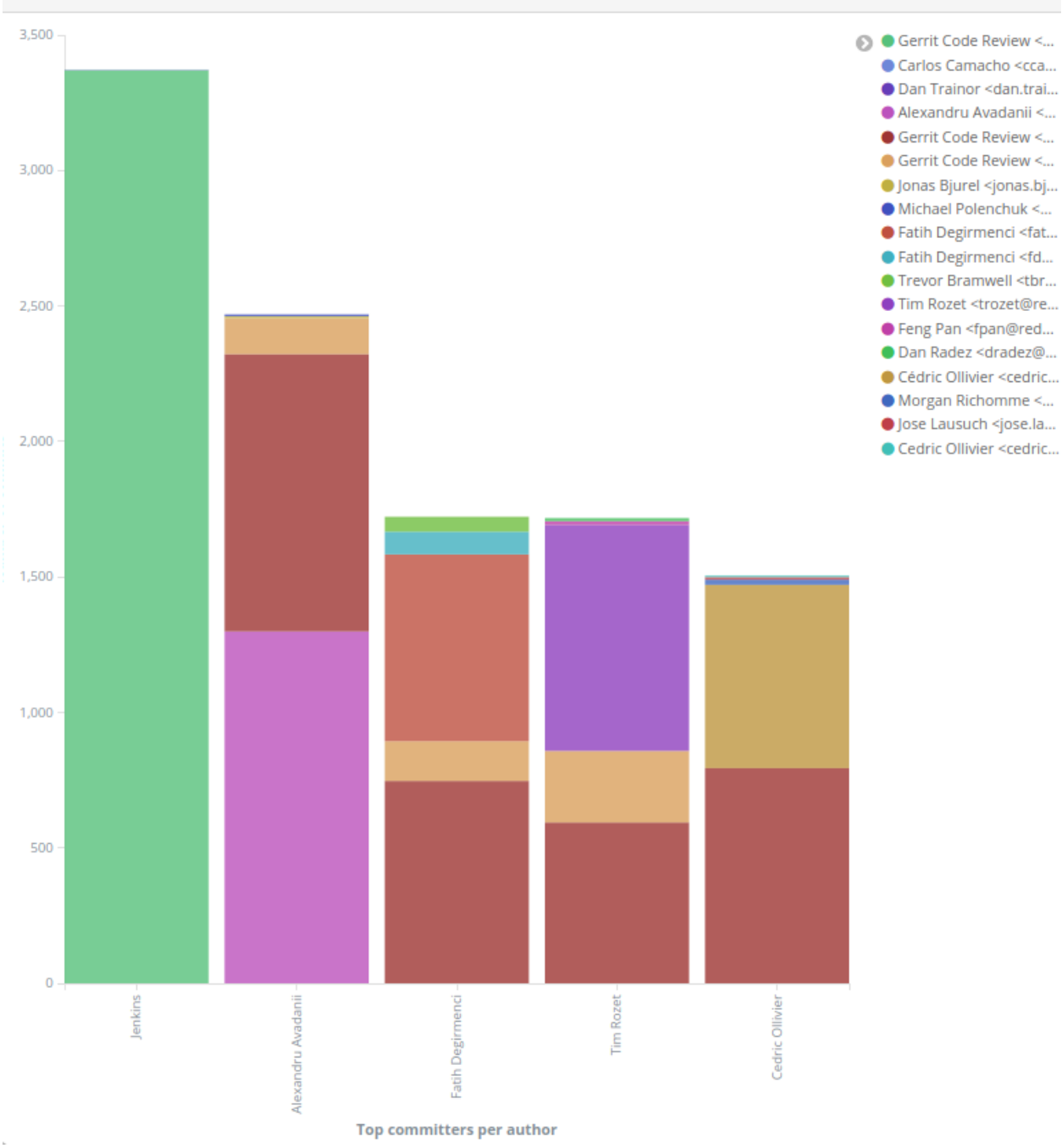
Organization	Author	Number
Ericsson	Fatih Degirmenci	1,746
Red Hat	Tim Rozet	1,730
Orange	Cedric Ollivier	1,519
Ericsson	Jose Lausuch	1,515
Orange	Morgan Richomme	1,483
Intel	Ross Brattain	1,117
Huawei	MatthewLi	682
Huawei	JingLu5	627
Red Hat	Dan Radez	597
Orange	Thomas Duval	494

En cuanto a las modificaciones, las personas que más modificaciones han hecho son (sumando el total de modificaciones hechas en cada commit):



Como se puede ver, no se corresponden las personas con mayor número de commits en comparación con las personas que hacen mayor número de modificaciones. Esto podría depender de los hábitos de las personas: algunos aprovechan un commit para hacer numerosos cambios y otras, en cambio, por suelen hacer commits para cambios pequeños.

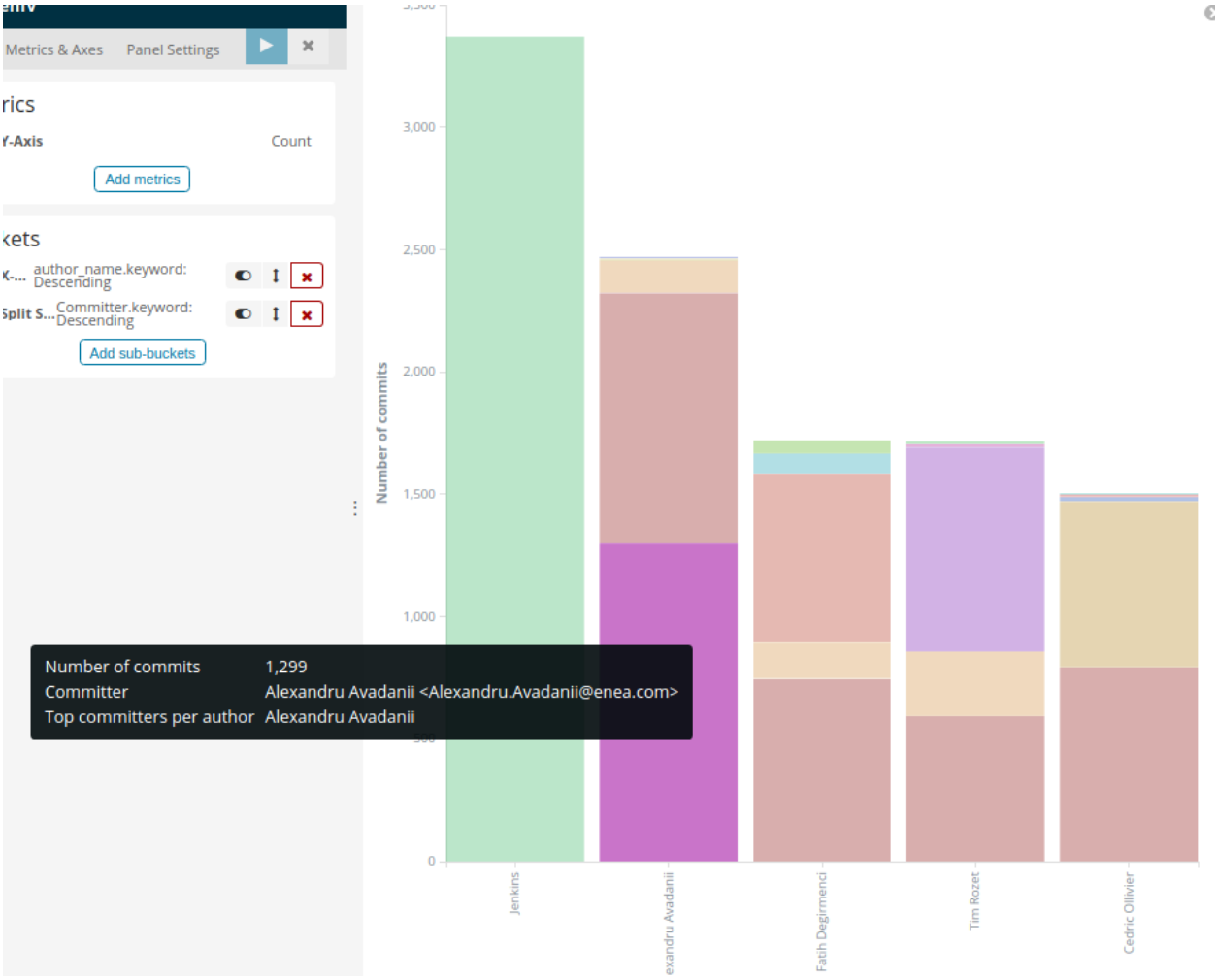
0.4.3 Committers per author



A continuación se muestra el listado de los mayores "committers" para cada autor:

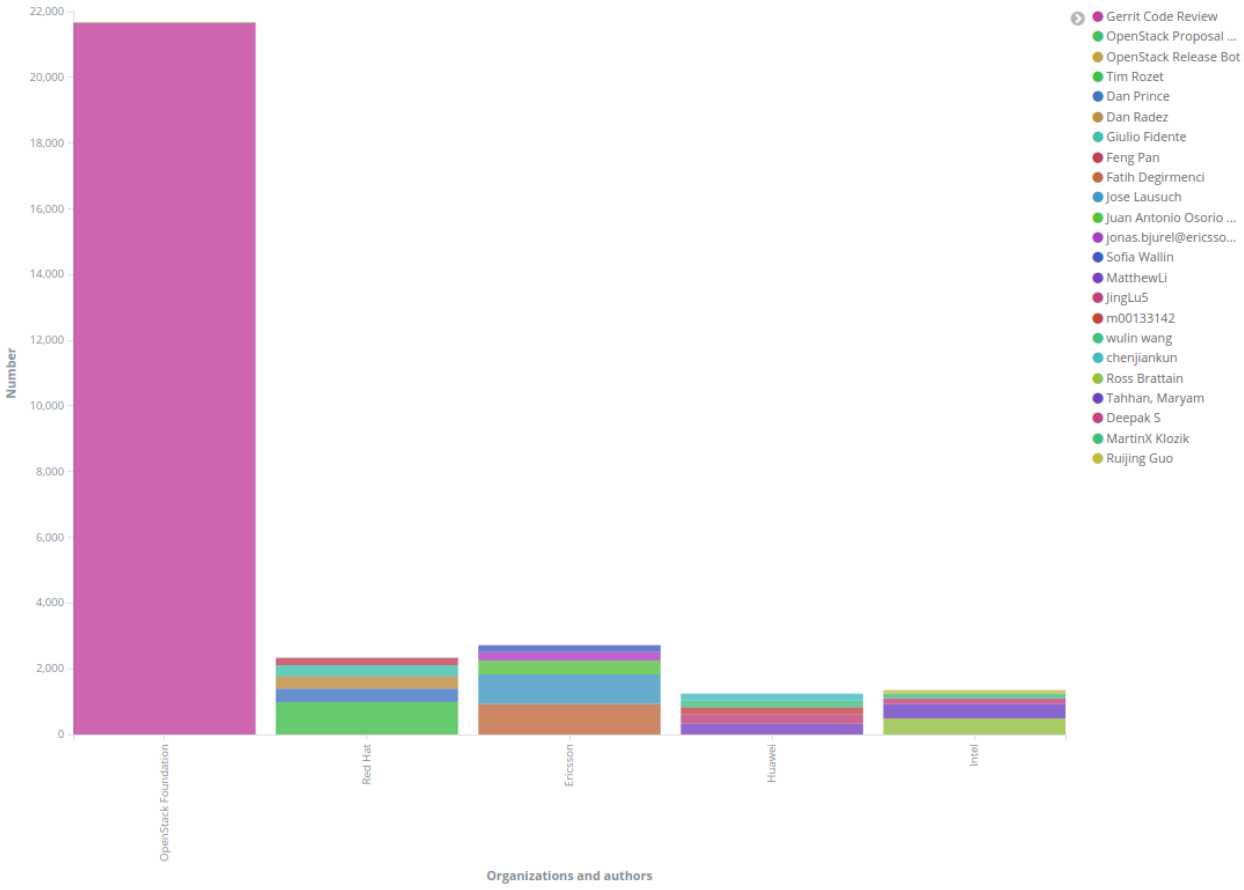
Author	Number	Committer	Number
Jenkins	3,372	Gerrit Code Review <review@openstack.org>	3,370
Jenkins	3,372	Carlos Camacho <ccamacho@redhat.com>	1
Jenkins	3,372	Dan Trainor <dan.trainor@gmail.com>	1
Alexandru Avadanii	2,480	Alexandru Avadanii <Alexandru.Avadanii@enea.com>	1,299
Alexandru Avadanii	2,480	Gerrit Code Review <gerrit@opnfv.org>	1,023
Alexandru Avadanii	2,480	Gerrit Code Review <gerrit@172.30.200.206>	133
Alexandru Avadanii	2,480	Jonas Bjurel <jonas.bjurel@ericsson.com>	7
Alexandru Avadanii	2,480	Michael Polenchuk <mpolenchuk@mirantis.com>	7
Fatih Degirmenci	1,746	Gerrit Code Review <gerrit@opnfv.org>	747
Fatih Degirmenci	1,746	Fatih Degirmenci <fatih.degirmenci@ericsson.com>	689

Es importante mencionar la importancia de Alexander Avadanii, pues es de los mayores autores, y cuenta con la suma de 1299. Por ello, podríamos pensar que tiene gran influencia.



0.4.4 Committer organizations

En lo que respecta a las "committer organizations" con mayor número de commits:



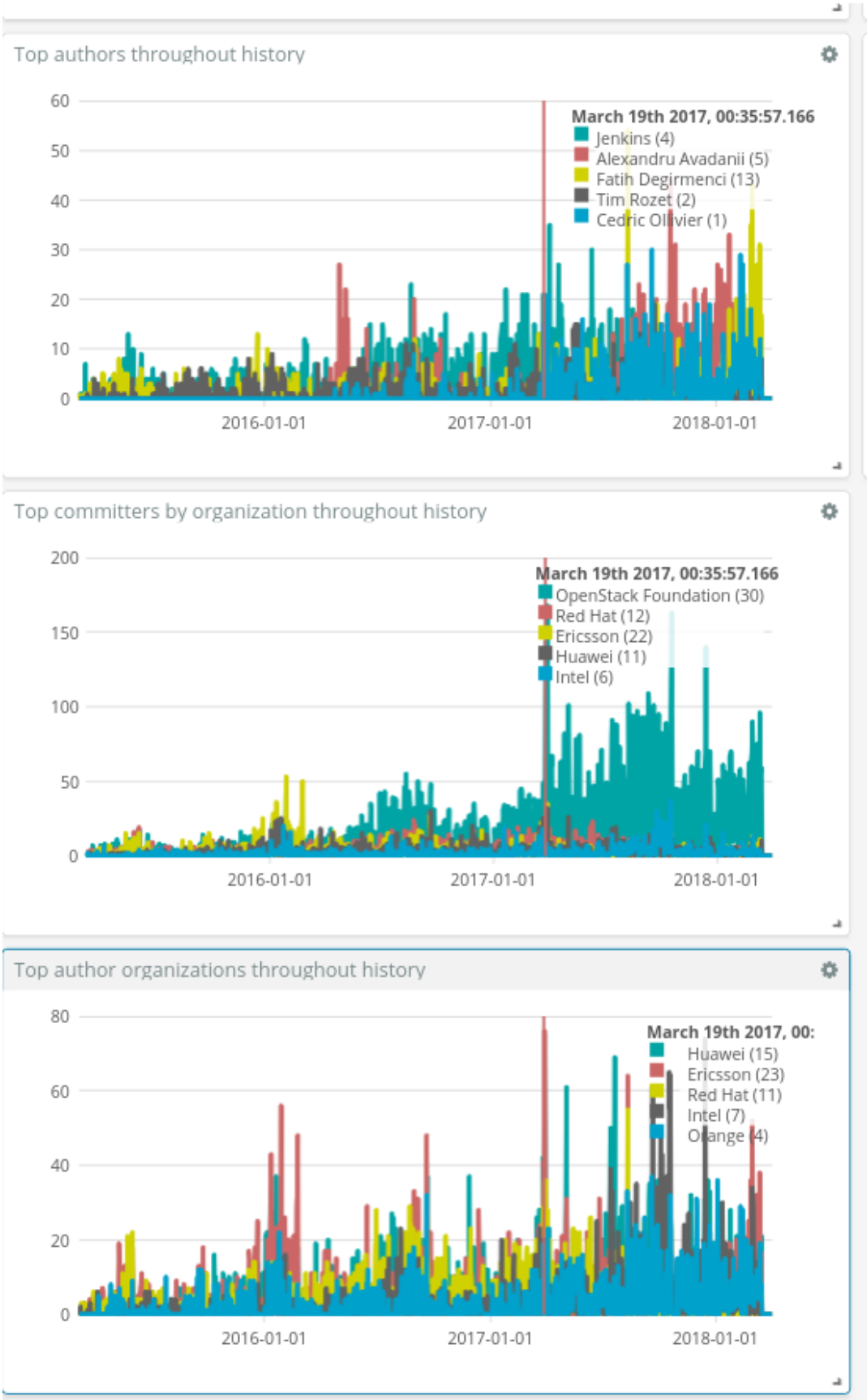
Como se puede ver, la "OpenStack Foundation" es, abrumadoramente, la que mayor número de commits ha hecho. Le siguen Ericsson, Red Hat, Intel y Huawei.

Committer organization	Committer	Number
OpenStack Foundation	Gerrit Code Review	21,661
Red Hat	Tim Rozet	988
Ericsson	Fatih Degirmenci	931
Ericsson	jose.lausuch	562
Intel	Ross Brattain	490
Intel	Maryam Tahhan	437
Ericsson	Juan Antonio Osorio Robles	414
Red Hat	Dan Prince	413
Red Hat	Dan Radez	363
Red Hat	Giulio Fidente	347

En lo que respecta a las personas que han realizado mayor número de commits según la compañía, para la "OpenStack Foundation" su líder es Gerrit Code Review, en Red Hat (Tim Rozet), en Ericsson (Faith Regirmenci), en Intel (Bross Brattain) y en Huawei (Linda Wang)

0.4.5 Evolución de la actividad a lo largo del tiempo

No hay ningún patrón de periodicidad en los datos a lo largo del tiempo, sin embargo, parece que todos los años, justo antes de su fin, hay un incremento considerable de la actividad. Además, hay que resaltar que en el año 2017 hubo un incremento muy notable de la actividad.


Top authors throughout history
60
50
40
30
20
10
0
March 19th 2017, 00:35:57.166
Jenkins (4)
Alexandru Avadanii (5)
Fatih Degirmenci (13)
Tim Rozet (2)
Cedric Ollivier (1)
2016-01-01
2017-01-01
2018-01-01
Top committers by organization throughout history
200
150
100
50
0
March 19th 2017, 00:35:57.166
OpenStack Foundation (30)
Red Hat (12)
Ericsson (22)
Huawei (11)
Intel (6)
2016-01-01
2017-01-01
2018-01-01
Top author organizations throughout history
80
60
40
20
0
March 19th 2017, 00:
Huawei (15)
Ericsson (23)
Red Hat (11)
Intel (7)
Orange (4)
2016-01-01
2017-01-01
2018-01-01

0.4.6 Conclusiones

La OpenStack Foundation es la organización que parece estar más comprometida con el proyecto de OPNFV. En cuanto a los autores, podemos ver que la tónica general es que desarrollen en una sola organización. Sin embargo, hay algún autor, como Faith Degirmenci que tiene gran actividad (tal y como hemos visto anteriormente) y colabora con proyectos de varios autores.

Además, podríamos mostrar algunos de los análisis que hemos llevado a cabo en el notebook.

0.4.6.1 Hábitos temporales (horas)

```
In [6]: #which are the utc hours when more files are changed.
        print(df.groupby(['hour'])['lines_changed'].mean())
        print(df.groupby(['hour'])['lines_removed'].mean())
        print(df.groupby(['hour'])['lines_added'].mean())
        #around mid -day in UTC it is when github register the most activity.
```

```
hour
00      81.320976
01    1039.629866
02     259.389374
03    1206.602542
04     639.579890
05     501.262411
06    3530.468182
07      50.081950
08    1328.465036
09     222.349542
10     615.606472
11     492.484245
12    1655.614265
13     352.388926
14     161.576066
15    5477.090465
16     171.829137
17     210.174572
18     166.851235
19      94.780048
20     125.008500
21     130.619699
22    1197.447809
23     378.726519
Name: lines_changed, dtype: float64
```

Como se puede ver, la mayor actividad se da a cabo sobre mediodía.

0.4.6.2 Hábitos temporales (mensual)

In [8]:
```
# which are the mnths when more files are changed.
print(df.groupby(['month'])['lines_changed'].sum())
print(df.groupby(['month'])['lines_removed'].sum())
print(df.groupby(['month'])['lines_added'].sum())
#around mid -day in UTC it is when github register the most activity.
#Surprisingly, August is the month when most lines are changed
```

```
month
a.January       1116031
b.February       477631
c.March          643827
d.April         3467135
e.May           5187496
f.June          1012042
g.July           602968
h.August       20070297
i.September     7728572
o.October       2043230
p.November      3565498
q.December       789303
Name: lines_changed, dtype: int64
month
a.January        714302
b.February       150373
c.March          389356
d.April          955097
e.May           4583955
f.June           361053
g.July           131101
h.August         133310
i.September     1355161
o.October        898436
p.November      1911321
q.December       193655
Name: lines_removed, dtype: int64
month
a.January        401729
b.February       327258
c.March          254471
d.April         2512038
e.May            603541
f.June           650989
g.July           471867
h.August       19936987
i.September     6373411
o.October       1144794
p.November      1654177
q.December       595648
Name: lines_added, dtype: int64
```

En lo que respecta a la actividad relativa a los meses sorprendentemente, en Agosto, es cuando se puede ver que hay mayor actividad.

0.4.6.3 Valoración de las empresas según las modificaciones realizadas

Podríamos hacer una valoración respecto a la compañía, teniendo en cuenta las modificaciones por cada commit.

	company	commits	lines_added	lines_removed	lines_changed
0	Huawei	6862	1883313.0	2599874.0	4483187.0
1	Ericsson	6304	405930.0	312100.0	718030.0
2	Red Hat	5971	249100.0	118630.0	367730.0
3	Intel	4270	25103233.0	4790527.0	29893760.0
4	Orange	3949	434512.0	252711.0	687223.0
5	ZTE Corporation	3576	313770.0	79804.0	393574.0
6	OpenStack Foundation	3457	16387.0	8856.0	25243.0
7	ENEA AB	2987	106342.0	98703.0	205045.0
8	Unknown	2076	3048466.0	1841038.0	4889504.0
9	Linux Foundation	1649	47884.0	176895.0	224779.0

Si comparásemos la tasa de modificaciones por cada commit, tendríamos la siguiente tabla:

	company	lines_added_per_commit	lines_removed_commit	lines_changed_commit	lines_removed_per_lines_changed
0	Huawei	274.455407	378.879918	653.335325	0.724386
1	Ericsson	64.392449	49.508249	113.900698	1.300641
2	Red Hat	41.718305	19.867694	61.585999	2.099806
3	Intel	5878.977283	1121.903279	7000.880562	5.240182
4	Orange	110.030894	63.993669	174.024563	1.719403
5	ZTE Corporation	87.743289	22.316555	110.059843	3.931758
6	OpenStack Foundation	4.740237	2.561759	7.301996	1.850384
7	ENEA AB	35.601607	33.044191	68.645798	1.077394
8	Unknown	1468.432563	886.819846	2355.252408	1.655841
9	Linux Foundation	29.038205	107.274106	136.312310	0.270692

Por un lado, la compañía RedHat parece un buen lugar en el que trabajar, puesto que es una empresa en la que se suelen dar pocos cambios por cada commit que se hace. Eso hace evidenciar que es una compañía con buenas prácticas profesionales. Más concretamente, tiene tan sélo 61 líneas cambiadas y un buen ratio de línea añadida por cada línea cambiada (esto significa que es una compañía en la que se suelen hacer pocos cambios por línea añadida) Por otro lado, Huawei es una compañía que realiza cambios enormes, muchos commits pero con gran número de líneas cambiadas por cada commit. De hecho, tiene una tasa de 0,72 líneas añadidas por cada línea cambiada, lo cual quiere decir que están continuamente refactorizando su código y puede que, por el volumen de trabajo, tenga gran número de proyectos abiertos.