ELASTICSEARCH Y KIBANA

Recuperación de Información y Minería de Texto

DESCRIPCIÓN BREVE

Analizar un gran volumen de datos de la comunidad OPNFV alojados en ElasticSearch, utilizando Kibana y la librería Python ElasticSearch DSL.

Remedios Blázquez Martin y Candela Vidal Rodríguez

Máster en Data Science 2017-2018

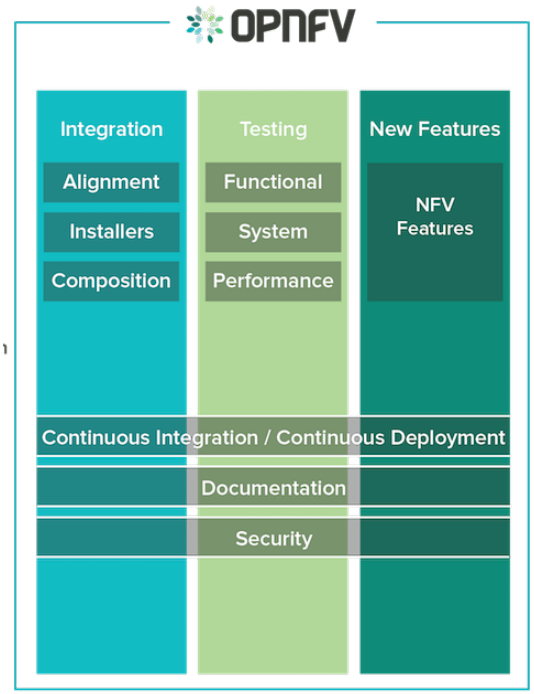
Índice

1. Introducción

La plataforma OPNFV es un proyecto abierto de la fundación Linux que facilita el desarrollo y evolución de los componentes de NFV (Network function virtualization).

EL proyecto OPNFV esta organizado alrededor de tres pilares claves:

- Integración: OPNFV integra una gran variedad de proyectos open source que cubren las necesidades específicas de NFV
- Testing: OPNFV testea el stack completo a través de parámetros como la funcionalidad, actuación y estrés.
- Nuevas características: La comunidad OPNFV desarrolla importantes características para varios de los proyectos open source integrados en ella. Todas estas características contribuyen a su vez en otros proyectos.



OPNFV se trata de una comunidad formada por miembros, desarrolladores, usuarios finales y miembros de comunidades de código abierto, y sus beneficios principales son:

- La integración de stack de redes de código abierto.
- El testing de arquitecturas específicas de referencia.
- Contribución integral de las características a nivel de operador.

2. Objetivos

A continuación, exponemos una breve declaración de las preguntas que buscamos resolver en esta práctica.

1. Resumen de organizaciones involucradas, así como desarrolladores y repositorios.
2. Principales organizaciones involucradas en Opnfv community, así como desarrolladores y los repositorios más activos.
3. Evolución a lo largo de estos últimos cinco años de las principales organizaciones, con respecto a los proyectos.
4. Identificar los desarrolladores que más contribuyen llevando a cabo un proyecto dentro de las principales organizaciones involucradas, y para cada uno de ellos sus repositorios más activos.
5. Evolución de los tres desarrolladores más activos por año, para ver si existe alguna relación con la compañía que en el año ha realizado más commits.
6. Evolución de commits por meses para ver si se detecta algún patrón significativo o no.

Previo análisis en profundidad de los datos, tanto con Kibana como con la librería Python ElasticSearch DSL, debemos aclarar cuál es la estructura de los datos que tenemos cargado en ElasticSearch.

Cada registro se corresponde con un commit realizado, y para cada commit tenemos:

- Fecha del autor y del commit
- Autor y committer
- Ficheros que se han tocado
- Líneas que se han tocado o eliminado
- Hash y mensaje

Los campos más relevantes dentro de todos los que nos ofrecen los datos, y que por tanto serán aquellos que utilicemos a la hora de realizar las consultas son:

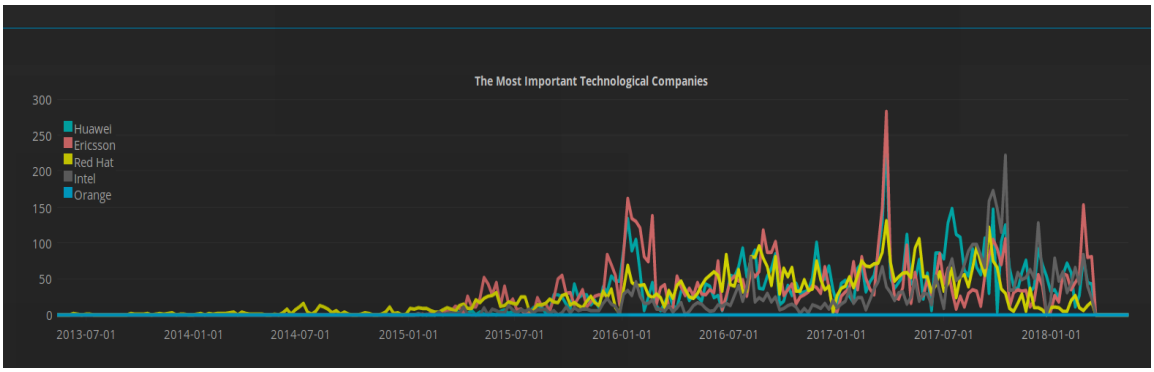
- grimoire_creation_date
- author_name
- author_org_name
- repo_name
- hash

3. Diseño del panel

Realizamos un Dashboard con diferentes visualizaciones para responder a las preguntas formuladas anteriormente.

I. Principales organizaciones o compañías involucradas en la comunidad Opnfv. Para ello la visualización elegida en de tipo "tag cloud", realizando un bucket por el campo: "author_org_name" y una métrica "count".

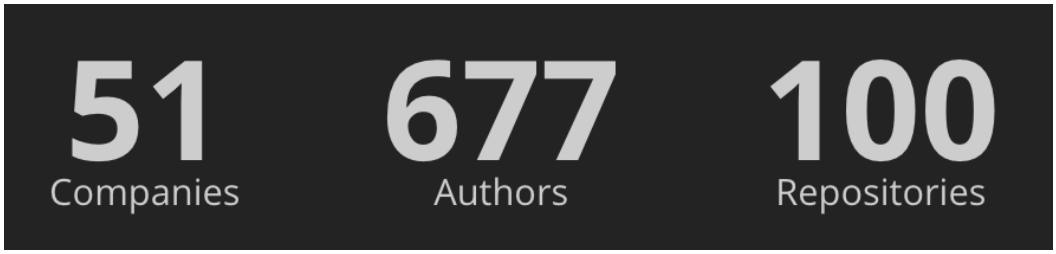
II. También hemos realizado una serie temporal por cada una de las compañías con respecto a los commits para ver su evolución en estos cinco años con la herramienta Timelion. Siendo una herramienta muy buena y de gran utilidad para comparar las diferentes time series.



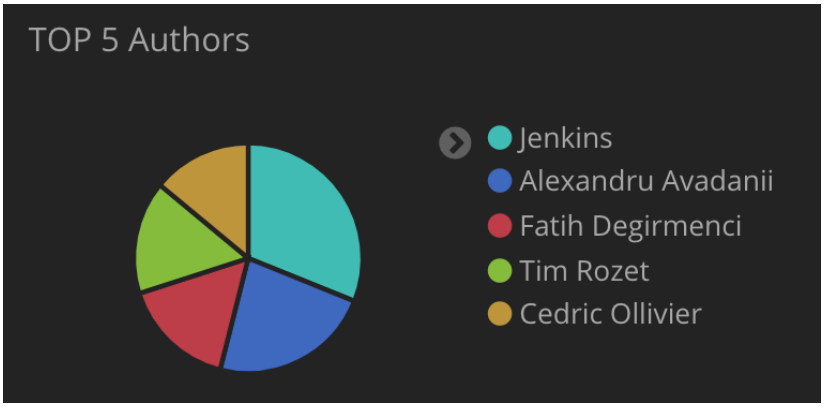
Timelion Expression

```
.es(q='Huawei', timefield='grimoire_creation_date',
index=git_opnfv_new).label(Huawei).title('The Most Important Technological Companies'),
.es(q='Ericsson', timefield='grimoire_creation_date',index=git_opnfv_new).label(Ericsson),
.es(q='Red Hat', timefield='grimoire_creation_date', index=git_opnfv_new).label('Red Hat'),
.es(q='Intel', timefield='grimoire_creation_date', index=git_opnfv_new).label(Intel),
.es(q='Organge', timefield='grimoire_creation_date', index=git_opnfv_new).label(Orange)
```

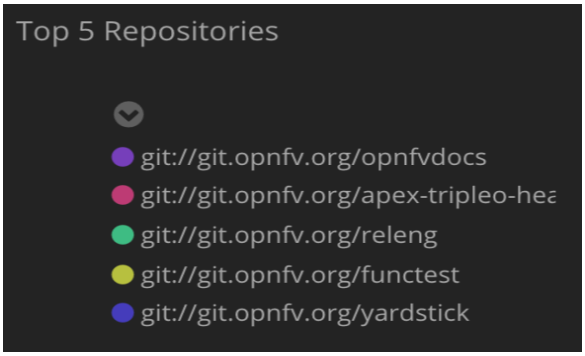
III. Summary, análisis cuantitativo de la totalidad de organizaciones involucradas en la comunidad Opnfv, así como la totalidad de desarrolladores y repositorios, visualización tipo "Metric".



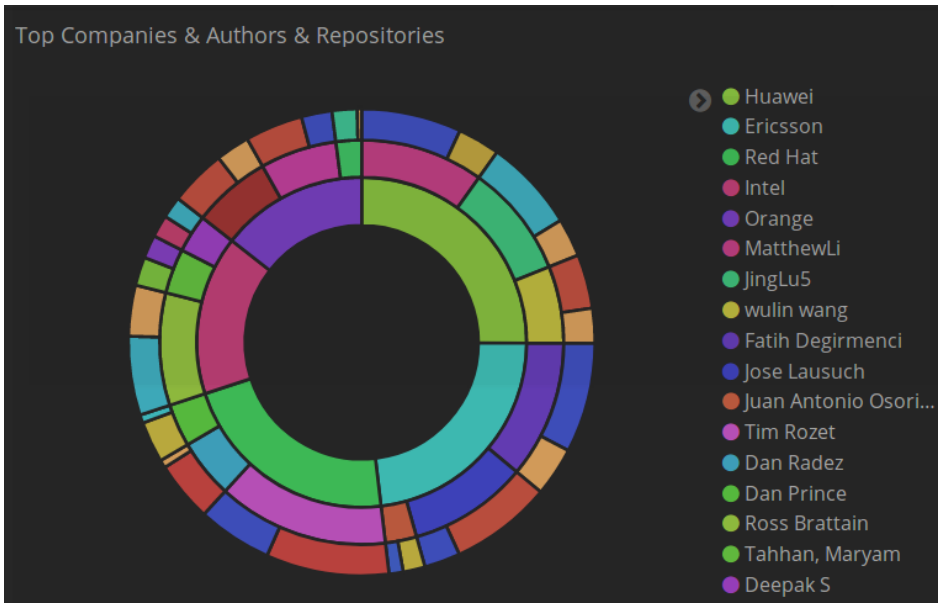
IV. Top 5 Authors, Top de los autores o desarrolladores que más participan en la comunidad opnfv durante todos los años. Visualización basada en un gráfico "Pie" normal, donde se muestra la leyenda en la parte derecha y hacemos uso del "tooltip". Se hace la métrica basada en "count" y un bucket por los términos del campo: "author_name".



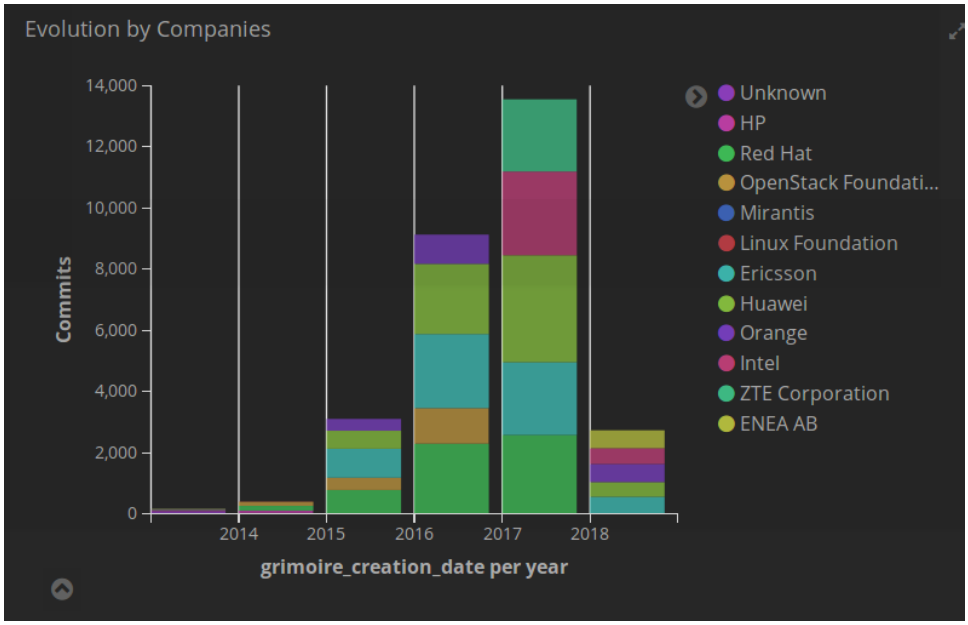
V. Top 5 Repositories, Top de los cinco repositorios más activos durante todos los años, hemos utilizado una visualización basada en un gráfico "Pie": tipo donut, y donde hacemos uso: "tooltip", ocultando la leyenda.



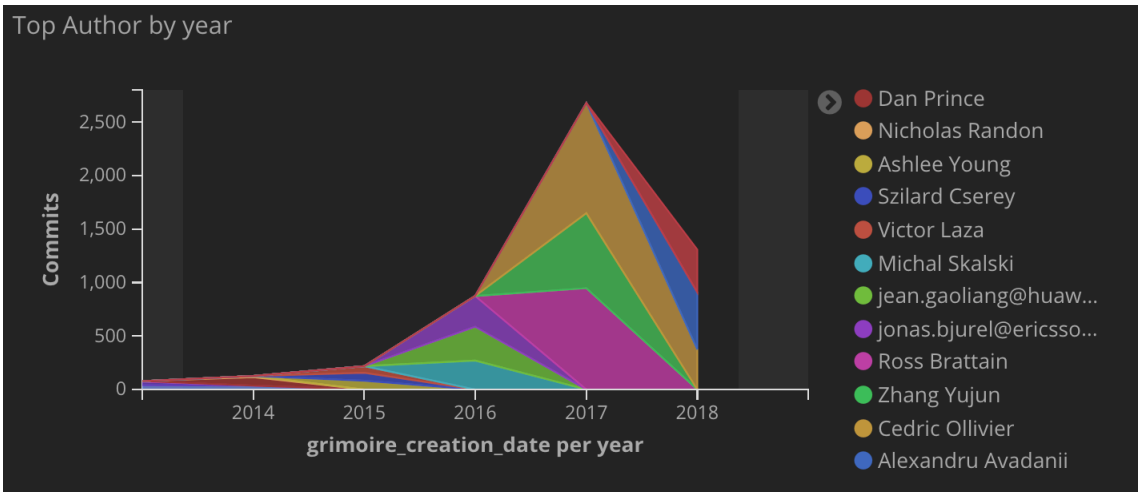
VI. Realizamos otra visualización para profundizar un poco más por cada una de las compañías más importantes, que desarrolladores contribuyen en cada una de ellas y por cada desarrollador los repositorios donde contribuyen más, hemos utilizado una visualización tipo "pie" donut, donde los datos se miden con la métrica "count" y realizamos diferente buckets, aquí queremos intentar ver si la nacionalidad de la organización o localización física está relacionada con la nacionalidad o cercanía continental de cada uno de los desarrolladores.



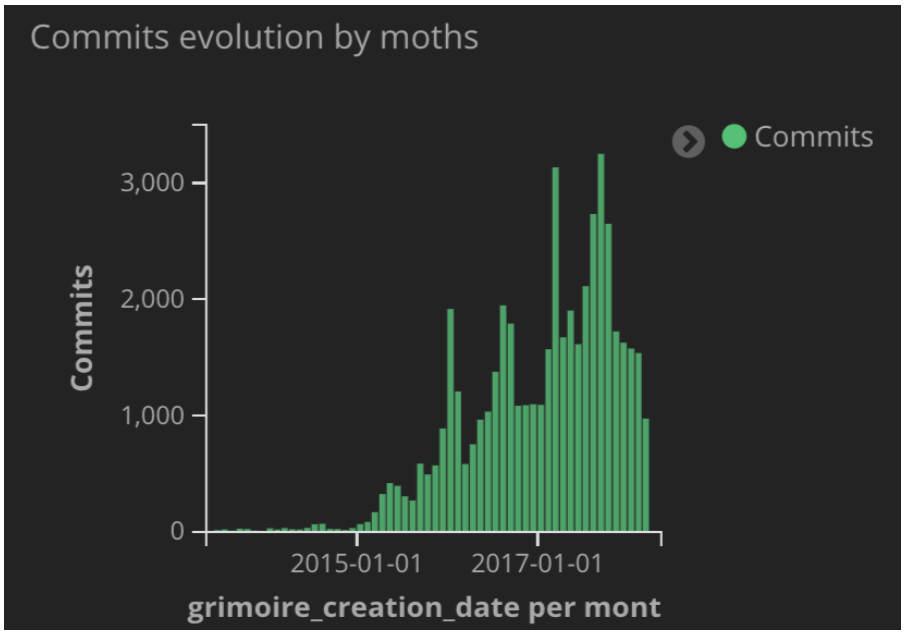
VII. Visualización de barras apiladas, donde la métrica utilizada es "count" y realizamos dos buckets por año y el campo "author_org_name" con un tamaño de 5 compañías por año. Aquí también detectamos otras compañías que compiten con las cinco grandes gigantes tecnológicas comentadas anteriormente.



VIII. Visualización basada en áreas apiladas, donde por cada año obtenemos los tres autores con más commits realizados, hacemos uso del "tooptip".



IX. Visualización de tipo barras, para ver la evolución del volumen de commits agrupados por mes.



4. Consultas del Notebook

Además del dashboard de Kibana, también hemos hecho uso de la librería de Python ElasticSearch DSL para realizar una serie de queries que soporten den respuesta a las preguntas planteadas en el apartado 2.

Nos hemos centrado en el análisis de las compañías, los autores y los repositorios a los que se han hecho los commits. De estos tres grupos hemos obtenido el número total, así como los 5 con mayor número de commits únicos:

Top 5 compañías:

```
s = Search(using=client, index=INDEX)
s.aggs.bucket('by_company', 'terms', field='author_org_name', size=5, order={'commits':'desc'}).metric('commits', 'cardinality', field='hash')
result = s.execute()
result.to_dict()['aggregations']

for company in result.to_dict()['aggregations']["by_company"]["buckets"]:
    print("company: " + company["key"])
    print('  commits: ' + str(company['commits']['value']))
```

```
company: Huawei
  commits: 6913
company: Ericsson
  commits: 6316
company: Red Hat
  commits: 5997
company: Intel
  commits: 4192
company: Orange
  commits: 3988
```

Top 5 autores:

```
s = Search(using=client, index=INDEX)
s.aggs.bucket('by_author', 'terms', field='author_name', size=5, order={'commits':'desc'}).metric('commits', 'cardinality', field='hash')
result = s.execute()
result.to_dict()['aggregations']

for author in result.to_dict()['aggregations']["by_author"]["buckets"]:
    print("author: " + author["key"])
    print('  commits: ' + str(author['commits']['value']))
```

```
author: Jenkins
  commits: 3363
author: Alexandru Avadanii
  commits: 2479
author: Fatih Degirmenci
  commits: 1745
author: Tim Rozet
  commits: 1730
author: Cedric Ollivier
  commits: 1519
```

Top 5 repositorios:

```
s = Search(using=client, index=INDEX)
s.aggs.bucket('by_repo', 'terms', field='repo_name', size=5, order={'commits':'desc'}).metric('commits', 'cardinality', field='hash')
result = s.execute()
result.to_dict()['aggregations']

for repo in result.to_dict()['aggregations']["by_repo"]["buckets"]:
    print("repo: " + repo["key"])
    print('  commits: ' + str(repo['commits']['value']))
```

```
repo: git://git.opnfv.org/opnfvdocs
  commits: 9712
repo: git://git.opnfv.org/apex-tripleo-heat-templates
  commits: 5938
repo: git://git.opnfv.org/releng
  commits: 5593
repo: git://git.opnfv.org/functest
  commits: 3880
repo: git://git.opnfv.org/yardstick
  commits: 2916
```

También de las 3 compañías con mayor número de commits hemos obtenido su top 3 de autores y los dos repositorios a los que más contribuyen cada uno de ellos:

```
s = Search(using=client, index=INDEX)
s.aggs.bucket('by_organization', 'terms', field='author_org_name',size=3, order={'commits':'desc'})\
.metric('commits', 'cardinality', field='hash')\
.bucket('by_author', 'terms', field='author_name', size=3, order={'commits':'desc'})\
.metric('commits', 'cardinality', field='hash')\
.bucket('by_repo', 'terms', field='repo_name', size=2, order={'commits':'desc'})\
.metric('commits', 'cardinality', field='hash')
result = s.execute()
result.to_dict()["aggregations"]

for key in result.to_dict()["aggregations"]["by_organization"]["buckets"]:
    print("company: " + key["key"])
    for author in key["by_author"]["buckets"]:
        print("  author: " + author["key"])
        for repo in author["by_repo"]["buckets"]:
            print("    repo: " + repo["key"])
```

Con el correspondiente resultado:

```
company: Huawei
  author: MatthewLi
    repo: git://git.opnfv.org/releng
    repo: git://git.opnfv.org/bottlenecks
  author: JingLu5
    repo: git://git.opnfv.org/yardstick
    repo: git://git.opnfv.org/opnfvdocs
  author: wulin wang
    repo: git://git.opnfv.org/functest
    repo: git://git.opnfv.org/opnfvdocs
company: Ericsson
  author: Fatih Degirmenci
    repo: git://git.opnfv.org/releng
    repo: git://git.opnfv.org/opnfvdocs
  author: Jose Lausuch
    repo: git://git.opnfv.org/functest
    repo: git://git.opnfv.org/releng
  author: Juan Antonio Osorio Robles
    repo: git://git.opnfv.org/apex-tripleo-heat-templates
    repo: git://git.opnfv.org/apex-puppet-tripleo
company: Red Hat
  author: Tim Rozet
    repo: git://git.opnfv.org/apex
    repo: git://git.opnfv.org/releng
  author: Dan Radez
    repo: git://git.opnfv.org/apex
    repo: git://git.opnfv.org/opnfvdocs
  author: Dan Prince
    repo: git://git.opnfv.org/apex-tripleo-heat-templates
    repo: git://git.opnfv.org/apex-os-net-config
```

Por último, hemos realizado una consulta para obtener los autores que habían trabajado para más de una compañía, intentando medir así de alguna manera el nivel de satisfacción de los desarrolladores dentro de la comunidad.

```
s = Search(using=client, index=INDEX)
s.aggs.bucket('by_author', 'terms', field='author_name', size=700)\
.bucket('by_org', 'terms', field='author_org_name')\
.metric('companies', 'cardinality', field='author_org_name')
result = s.execute()
orgs = result.to_dict()["aggregations"]["by_author"]["buckets"]
for org in orgs:
    if len(org["by_org"]["buckets"])!=1:
        print("author: " + org['key'])
        for company in org["by_org"]["buckets"]:
            print("  company: " + company["key"])
```

```
author: Nauman_Ahad
  company: Dell
  company: Unknown
author: baigk
  company: Huawei
  company: Unknown
author: shiva-charan.m-s
  company: Hewlett Packard Enterprise Co.
  company: HP
```

5. Análisis de los datos y conclusiones obtenidas.

A) En la visualización inicial (I) sobre el "Top compañías" apreciamos un ranking donde en primer lugar se sitúa: Huawei (25.08%), segundo lugar: Ericsson (23.04%) y tercer lugar: Red Hat (21.83%), seguidas de Intel (15.61%) y Orange (14.44%), con esto podemos contemplar que las compañías Europeas en las que se encuentran: "Ericsson y Orange" abarcan un 37,48% de la totalidad, muy seguidas de las compañías Americanas que acumulan un 37,44% y de la Asiática "Huawei" con un 25.08%, esto nos lleva a pensar de que manera se distribuyen los proyectos de la comunidad Opnfv por todo el mundo, sobre todo por los países más desarrollados del planeta, situándose en espacios continentales bien diferenciados.

B) Las series temporales graficadas con Timelion nos ayuda a ver la evolución de cada una de las grandes tecnológicas:
   a. Huawei, compañía emergente en el año 2016, con fuerte pendiente positiva manteniéndose con grandes altibajos durante un periodo de un año, apreciándose un descenso de volumen de commits en el 2018.
   b. Ericsson, compañía con importante repunte a mediados del 2015, se observa patrón de estacionalidad, más estable a lo largo de los años.
   c. Red Hat, su evolución se ve consolidada a lo largo del tiempo, a pesar de que en el último año se aprecia una disminución, su carrera en la colaboración con la comunidad opnfv es iniciada en el 2014.

C) Con las visualizaciones (VI, VII, VIII) queremos detectar si alguno de los desarrolladores tiene alguna relación o vinculo con alguna compañía.

   Observamos que para "Huawei", sus desarrolladores son todos o al menos su identificación es de origen asiático, mientras que para "Ericsson" son de origen europeo: "Fatih Degirmenci", "Jose Lausuch","Juan Antonio Osorio Robles".

   Siendo Ericsson la compañía en el año 2015, 2016 con más commits realizados, el autor con más commit para el año 2016 pertenece a dicha compañía: jonas.bjurel@ericsson.com, y el año 2015: "Víctor Laza".

   Todo esto nos lleva a pensar que hay cierto vinculo que pueda ser la nacionalidad o la cercanía continental lo que vincula compañías con desarrolladores.

También detectamos otras compañías que, no perteneciendo al top de las cinco compañías, aparecen en esta evolución como:

   a. "OpenStack Foundation" observando una creciente evolución desde 2014 al 2016 y sin más no obtener más datos de ella, ¿Qué ha ocurrido con esta compañía, sigue estando en opfnv activa, al igual que "HP", “Mirantis”, ¿”Linux Foundation”?

b. También nos hace pensar que quizás a partir del año 2017, emergen nuevas compañías como: "ZTE corporation", "ENEA AB", "Intel", ya que ninguna de ellas aparece en años anteriores.
c. ¿Mientras que las “compañías top” su evolución se prologa en los diferentes años, a que se debe?

Con la visualización IX, hemos detectado cierta estacionalidad a la hora de llevar a cabo los proyectos, observando que en el tercer trimestre del año es cuando se producen más commits.

¿A que se debe? ¿Los desarrolladores tiene más tiempo libre en ese periodo y pueden colaborar más o simplemente es puro azar? Para resolver estas preguntas deberíamos seguir investigando en futuros trabajos...

En conclusión:

"Grandes compañías multinacionales tecnológicas colaboran con la comunidad opnfv donde trabajan desarrolladores de todos los continentes colaborando en multitud de proyectos de software abierto"