Práctica – Recuperación de Información: Elasticsearch & Kibana

Autores

Ignacio Arias Barra

Raúl Sánchez Martín

Índice

1 Objetivo, documentación aportada y estructura de la entrega

El objetivo de la presente práctica es analizar la comunidad OPNFV (Open Platform for NVF [1]) utilizando conjuntamente Elasticsearch, Kibana y Python. Para ello, serán utilizados los datos proporcionados durante la docencia de la asignatura de Recuperación de Información del Máster en Data Science de la Universidad Rey Juan Carlos. A continuación se detalla toda la documentación aportada en la presente entrega:

- La presente memoria en formato *docx* y *PDF* (*Memorial_IgnacioArias_RaulSanchez.docx* y *Memorial_IgnacioArias_RaulSanchez.pdf*).
- Un Notebook (*Notebook_IgnacioArias_RaulSanchez.ipynb*) en el que se recoge, de manera ordenada y detallada, todo el código desarrollado en Python para la ejecución y posterior visualización de las diferentes consultas propuestas.
- Todas las visualizaciones (carpeta *visualizaciones_python*), en formato *png*, obtenidas a partir del anterior Notebook. Cabe aclarar que dichas visualizaciones están incluidas en el propio Notebook, pero de manera adicional se han copiado en una carpeta independiente para facilitar su uso. La información de cada consulta está organizada en diferentes subcarpetas. En cada una de ellas, además de las figura en formato *png*, también se incluye un *json* que recoge los resultados numéricos de dichas figuras.
- Todas las visualizaciones (carpeta *visualizaciones_kibana*), en formato *png* y *json*, obtenidas a partir de Kibana. Esta información se ha dividido en dos carpetas:
    - *dashboard*: En dicha carpeta, se incluye toda la documentación necesaria para la visualización del dashboard realizado en esta práctica en Kibana. Por un lado, el fichero *dashboard.json* necesario para su visualización en Kibana. Por otro lado, en la subcarpeta *visualizaciones_individuales*, se incluyen las diferentes figuras en formato *png* que forman el dashboard.
    - *consultas_individuales*: En dicha carpeta, se incluyen tantas subcarpetas como consultas propuestas. Y por cada subcarpeta, diferentes figuras en formato *png* que han sido obtenidas en Kibana, incluyendo resultados tanto gráficos como numéricos.

El presente documento se estructura como sigue. En primer lugar, se incluye una introducción a la comunidad OPNFV. Posteriormente, se describen las diferentes consultas propuestas sobre dicha comunidad. A continuación, se describen los diferentes resultados obtenidos para cada consulta. Finalmente, se incluyen las principales conclusiones obtenidas del presente estudio.

2 Introducción a OPNFV

¿Qué es?

El proyecto OPNFV consiste en una plataforma que facilita la innovación de la virtualización de funciones de red. Dicha plataforma es de carácter *open source*, lo que facilita el acceso y uso de la misma. Esta nueva forma de interoperabilidad entre desarrolladores, empresas de infrastructura, cloud y clientes, promoverá la creación de nuevos servicios y acelerará el desarrollo de funcionalidades NFV (Network Functions Virtualizaion).

¿A qué se dedica?

OPNFV contiene la aportación de código de todo desarrollador y empresa que quiera introducirse en esta plataforma. Esto facilita la unificación de trabajos y acelera el desarrollo y la rápida integración continua de servicios. Además, la comunidad es capaz de llevar a cabo la monitorización del rendimiento y el uso de pruebas basadas en diversas soluciones para asegurar la idoneidad de la plataforma para casos de uso NFV.

OPNFV trabaja junto con otras comunidades *open source* de tal forma que se aúnan contribuciones de código y nuevas soluciones. En esta comunidad se están incluyendo

continuamente diversos desarrolladores que aportan diferentes ideas, diferentes tipos de conocimiento y haciendo que los proyectos se desarrollen en un menor tiempo y con soluciones más robustas.

¿Qué pretende conseguir?

La plataforma tiene como objetivo facilitar y acelerar el desarrollo en conjunto y la consolidación de soluciones NFV a través de diversos ecosistemas de código abierto. Gracias a la integración, despliegue y pruebas continuas, se consigue una plataforma NFV referente para empresas y proveedores de servicios. La participación está abierta a cualquiera, ya sea empresa, desarrollador independiente u organización.

¿Qué datos manejan?

OPNFV está formada por todo el código aportado en la plataforma tanto por desarrolladores como por empresas.

¿Con qué datos vamos a trabajar nosotros?

En la presente práctica se realizará un análisis de la actividad de aportación de código de un gran número de desarrolladores de la comunidad OPNFV. Los datos que se manejarán consisten en una base de datos cuyos donde se incluyen los *commits* que los diferentes desarrolladores han realizado. Algunos campos interesantes de los datos a analizar son *nombre del desarollador*, *empresa a la que pertenece*, *id del commit*, etc.

3 Consultas propuestas

A continuación se van a describir las diferentes consultas que se han propuesto sobre la base de datos estudiada para analizar así el comportamiento de la comunidad OPNFV. En concreto, se proponen las 6 consultas siguientes:

- *Consulta 1:* ¿Cuáles son las principales organizaciones que participan?
- *Consulta 2:* ¿Cuáles son los principales desarrolladores que toman parte de la comunidad?
- *Consulta 3:* ¿Cuáles son los repositorios más activos?
- *Consulta 4:* ¿Cuáles son las organizaciones con más proyectos?
- *Consulta 5:* Se analizará la actividad de una empresa en concreto que será elegida en función de los resultados obtenidos en las preguntas anteriores.
- *Consulta 6:* ¿Cuáles son los proyectos más importantes?

4 Resultados

A continuación se van a incluir las diferentes soluciones propuestas para responder a las consultas descritas anteriormente. Cada una de las consultas ha sido resuelta por medio de diversas visualizaciones tanto en Kibana como en Python. Los resultados de cada consulta solucionada en Kibana se incluyen en la carpeta *visualizaciones_kibana/consultas_individuales*. Por otro lado, las mismas visualizaciones también han sido resueltas en Python. Para ello, en primer lugar se ha realizado una query a la misma base de datos utilizando el cliente de Python preparado para Elasticsearch. Los resultados obtenidos han sido guardados en un *json*. Posteriormente, dichos datos se han visualizado utilizando Pandas y Matplotlib. Todo el código necesario para resolver cada consulta está incluido en el Notebook. Además, los resultados extraídos de esta manera (tanto la figura como el *json*), están recogidos en la carpeta *visualizaciones_python*.

En la siguiente sección, se van a mostrar y comentar los resultados de cada una de las consultas utilizando las gráficas obtenidas a partir de Python. A continuación, se mostrará un dashboard equivalente al que hemos realizado para Kibana, incluyendo las figuras individuales obtenidas por este software.

4.1 Solución individual de cada consulta

4.1.1 Consulta 1: ¿Cuáles son las principales organizaciones que participan?

Para responder a esta consulta, vamos a analizar dos aspectos. Por un lado, el número de commits realizados por los miembros de las diferentes organizaciones. Y por otro, el número de desarrolladores que pertenecen a cada organización. Nos vamos a centrar en los 10 primeros resultados en cada caso. Ambos aspectos son mostrados en Fig. 1 y Fig. 2 respectivamente.

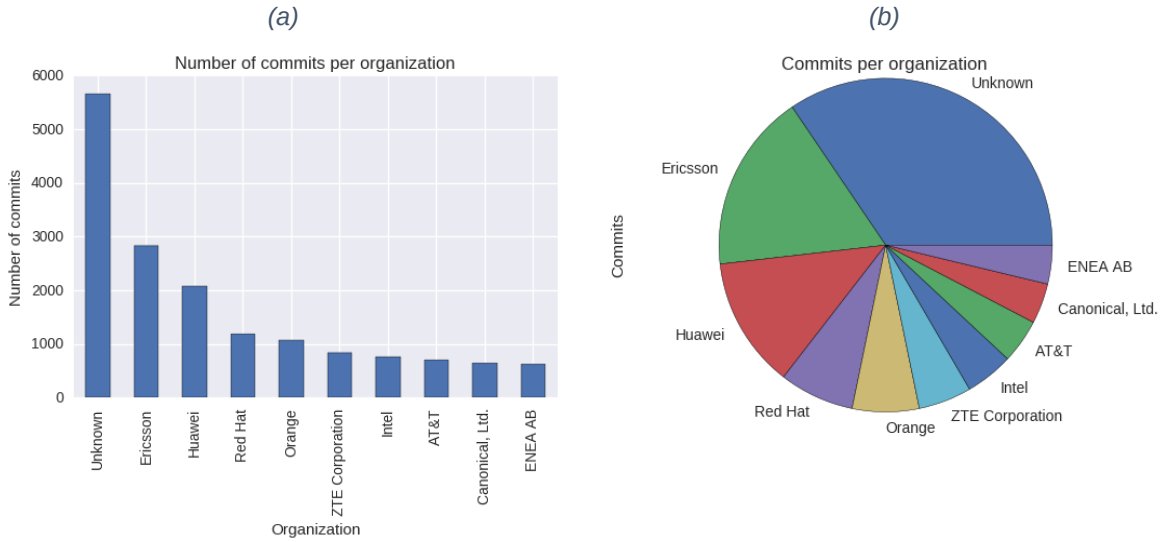


Fig. 1: Número de commits por organización: (a) Barplot; (b) Piechart.

Código python query:

```
s = Search(using=client, index=INDEX).filter('range', author_date={'gt': datetime(2015, 1, 1)})
s.aggs.bucket('by_company', 'terms', field='Commit_org_name').metric('Commits', 'value_count', field='Commit_id')
```

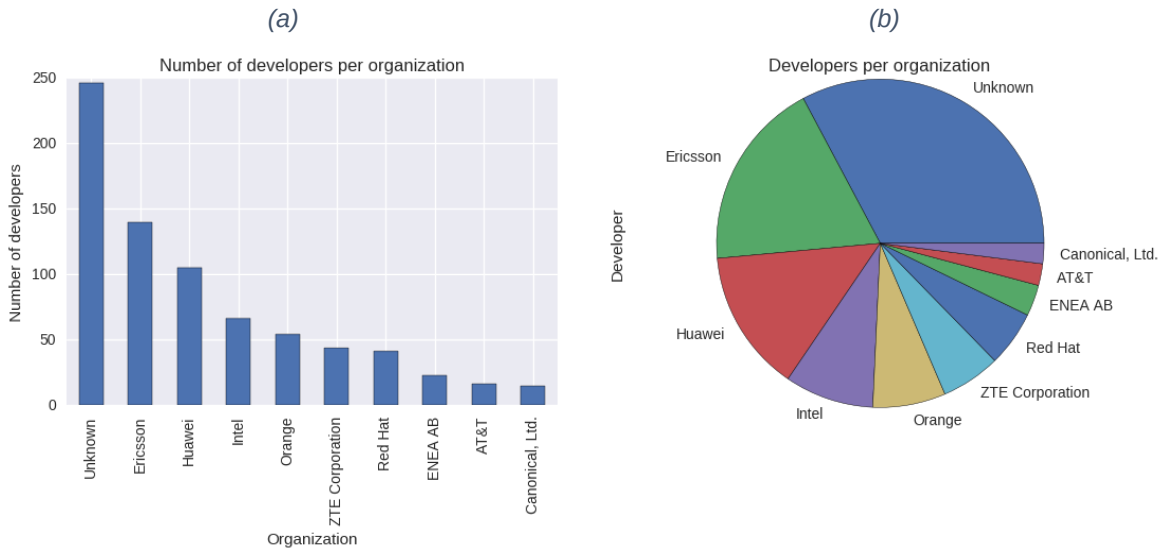


Fig. 2: Número de desarrolladores por organización: (a) Barplot; (b) Piechart

Código python query:

```
s = Search(using=client, index=INDEX).filter('range', author_date={'gt': datetime(2015, 1, 1)})
s.aggs.bucket('by_company', 'terms', field='Commit_org_name').metric('Developer', 'cardinality', field='Author')
```

Como se puede observar, utilizando ambas métricas, las dos organizaciones que más participan son (excluyendo aquellos commits/desarrolladores que no pertenecen a ninguna), y por este orden, Ericsson y Huawei. Sin embargo, utilizando la métrica de los commits, la tercera organización es Red Hat, organización que cae hasta el 6 puesto en cuanto a número de desarrolladores. Eso podría indicar que esta organización tiene desarrolladores, que aunque no muy numerosos, realizan una gran cantidad de commits. Una lectura complementaria se podría realizar de Intel, que ocupa el tercer puesto en cuanto a desarrolladores, pero el sexto en cuanto a número de commits.

4.1.2 Consulta 2: ¿Cuáles son los principales desarrolladores que toman parte de la comunidad?

Para estudiar cuáles son los desarrolladores principales, vamos a analizar por un lado el número de commits realizado por cada uno de ellos, y por otro el número de líneas añadido por cada uno de ellos. Ambos aspectos están incluidos en la Fig. 3 y Fig. 4.

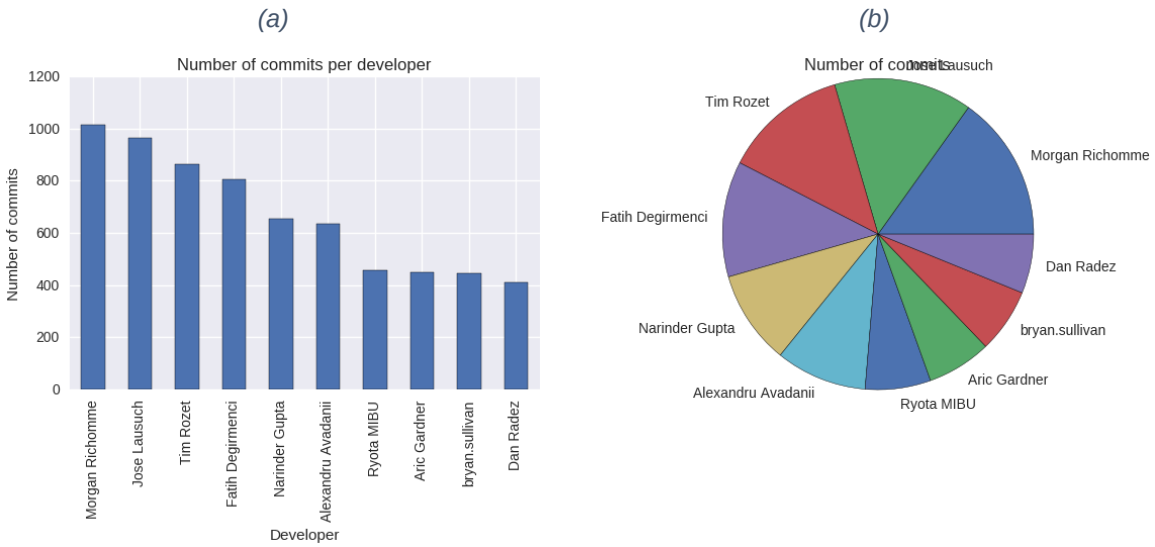


Fig. 3: Número de commits por desarrollador: (a) Barplot; (b) Piechart.

Código python query:

```
s = Search(using=client, index=INDEX).filter('range', author_date={'gt': datetime(2015, 1, 1)})
s.aggs.bucket('by_author', 'terms', field='author_name').metric('commits_per_developer', 'cardinality', field='hash')
```

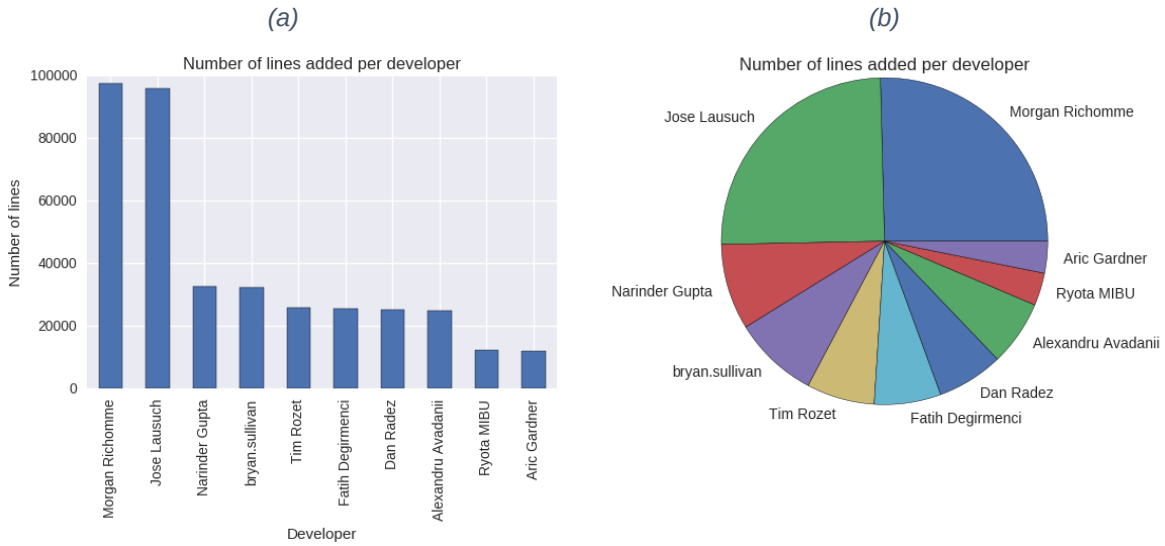


Fig. 4: Número de líneas añadidas por desarrollador: (a) Barplot; (b) Piechart.

Código python query:

s = Search(using=client, index=INDEX).filter('range', author_date={'gt': datetime(2015, 1, 1)})
s.aggs.bucket('by_author', 'terms', field='author_name').metric('lines_added_per_developer', 'sum', field='lines_added')

Utilizando ambas métricas, los principales desarrolladores son, por este orden, Morgan Richomme y Jose Lausuch. Sin embargo, Tim Rozet aparece en el tercer puesto en cuanto al número de commits, y sin embargo cae al puesto 5 en función de las líneas de código añadidas. Esto puede representar que cada uno de sus commits incluye, de media, menos líneas de código que sus compañeros. La lectura contraría se puede realizar de Narinder Gupta, que ocupa la quinta posición en cuanto a número de commits pero la tercera respecto al número de líneas añadidas.

4.1.3 Consulta 3: ¿Cuáles son los repositorios más activos?

Para analizar cuáles son los repositorios más activos, vamos a seguir el mismo esquema que hemos utilizado para analizar las organizaciones más importantes. Es decir, se van a contabilizar, por cada proyecto, por un lado, el número de commits registrados, y por otro, el número de desarrolladores que participan. Ambos aspectos vienen recogidos en la Fig. 5 y Fig. 6 respectivamente.

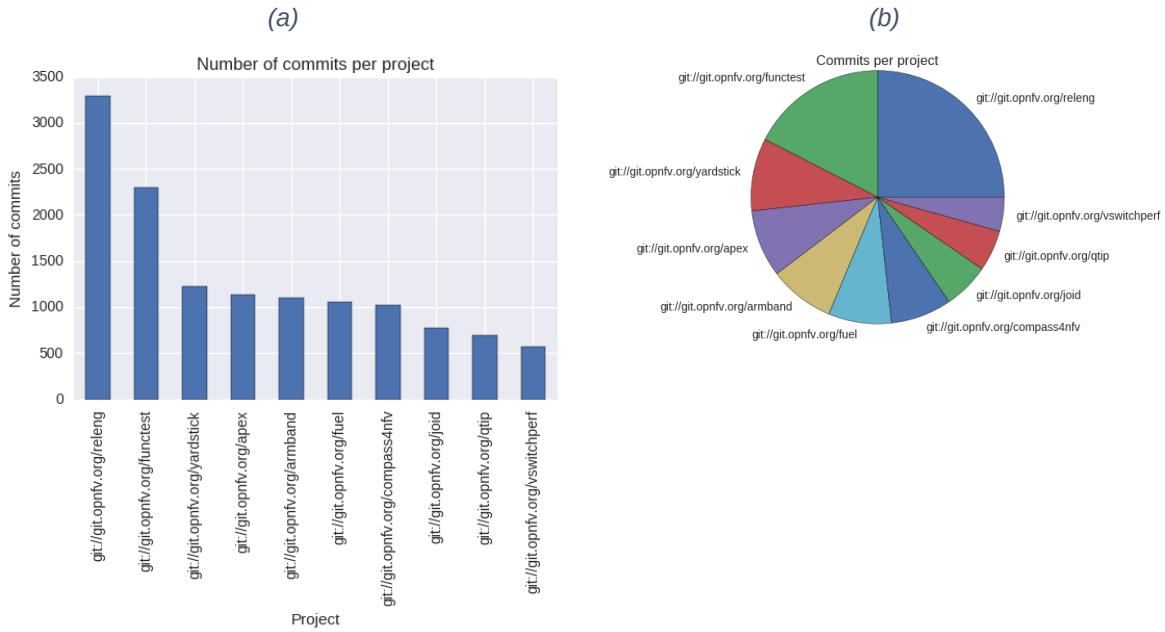


Fig. 5: Número de commits por proyecto: (a) Barplot; (b) Piechart.

Código python query:

s = Search(using=client, index=INDEX).filter('range', author_date={'gt': datetime(2015, 1, 1)})
s.aggs.bucket('by_repo', 'terms', field='repo_name').metric('Commits', 'value_count', field='Commit_id')

Como se puede observar, los dos principales proyectos son, con diferencia, releng y functest. Posteriormente, hay un conjunto de diferentes proyectos (yardstick, armband, fuel, compass4nfv, etc…) que están bastante igualados.

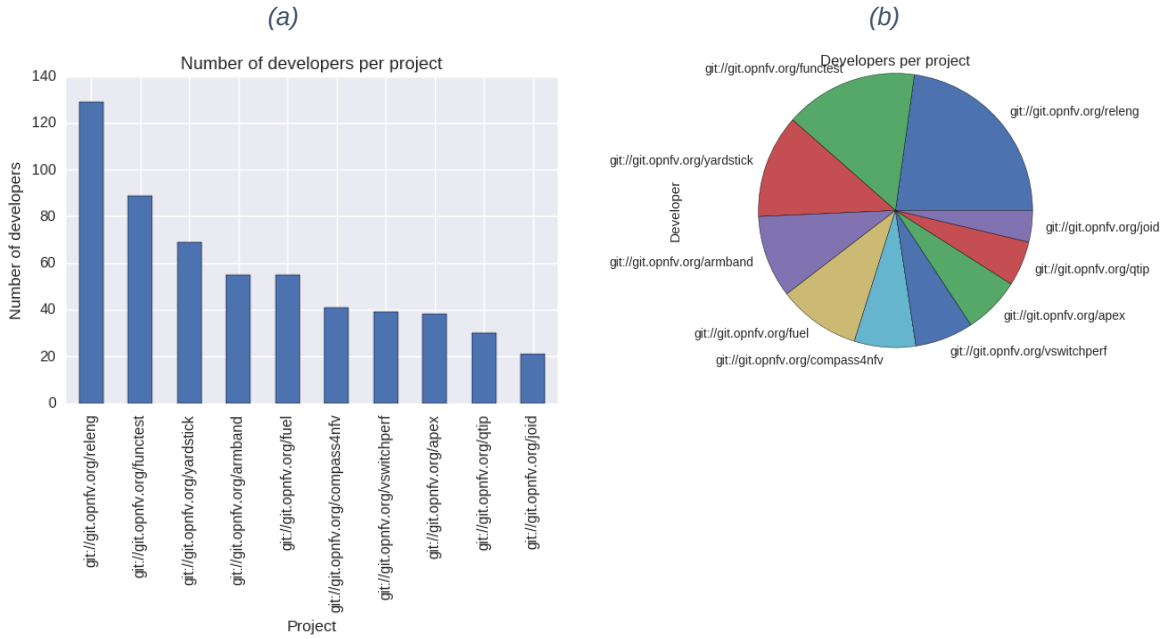


Fig. 6: Número de desarrolladores por proyecto: (a) Barplot; (b) Piechart.

Código python query:

s = Search(using=client, index=INDEX).filter('range', author_date={'gt': datetime(2015, 1, 1)})
s.aggs.bucket('by_repo', 'terms', field='repo_name').metric('Developer', 'cardinality', field='Author')

4.1.4 Consulta 4: ¿Cuáles son las organizaciones con más proyectos?

La presente consulta es complementaria a la primera. En este caso, también se mide la importancia de cada organización, pero no por el número de commits o desarrolladores, si no por el número de proyectos en los cuales están involucrados. Este aspecto viene representado en la Fig. 7.

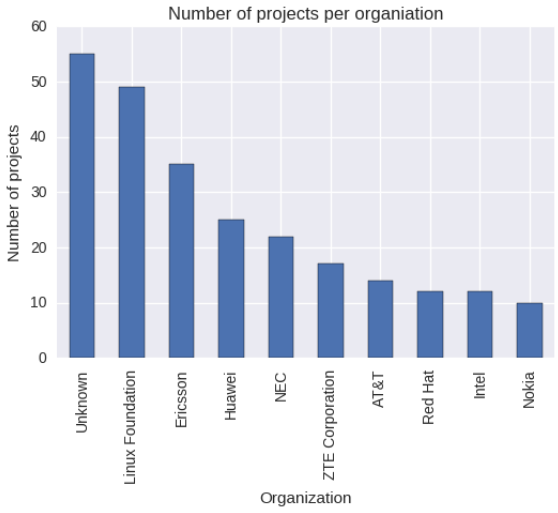


Fig. 7: Número de proyectos por organización.

Código python query:

```
s = Search(using=client, index=INDEX).filter('range', author_date={'gt': datetime(2015, 1, 1)})
s.aggs.bucket('by_company', 'terms', field='Commit_org_name', size=50).metric('Projects', 'cardinality', field='repo_name')
```

Como se puede observar, Ericsson y Huawei siguen siendo muy importantes en este sentido. Sin embargo, una organización que no aparecía en la primera consulta, Linux Foundation, ahora acapara la primera posición. Esto puede deberse a que tiene un gran número de proyectos los cuales acumulan de manera individual poca actividad.

4.1.5 Consulta 5: Se analizará la actividad de una empresa en concreto que será elegida en función de los resultados obtenidos en las preguntas anteriores.

Para responder a esta pregunta, vamos a elegir la empresa Ericsson, debido a su importancia de acuerdo a las consultas anteriores. En concreto, vamos a analizar el número de commits realizados por desarrolladores de esta empresa a lo largo del tiempo (Fig. 8, conteo semanal).

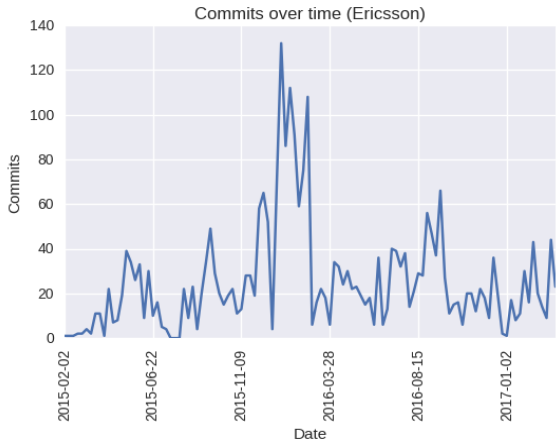


Fig. 8: Evolución de los commits para la empresa Ericsson.

Código python query:

```
company = "Ericsson"
s = Search(using=client, index=INDEX).filter('range', author_date={'gt': datetime(2015, 1, 1)})
s = s.filter("term",Commit_org_name=company)
s.aggs.bucket('histogram', 'date_histogram', field='commit_date', interval='week').metric('commits', 'cardinality', field='hash')
```

Como se puede observar, la aportación de los desarrolladores de la empresa Ericsson a la comunidad OPNVF ha sido más o menos regular menos en un periodo de unos 3-4 meses, a finales del año 2015 y principios del 2016, donde su actividad se multiplicó casi por tres. Esto puede ser debido a la necesidad de acabar un gran proyecto por parte de esta empresa durante este periodo.

4.1.6 Consulta 6: ¿Cuáles son los proyectos más importantes por cada organización?

En consultas anteriores, ya se ha analizado cuales son los proyectos más importantes. Pero no se ha hecho distinción si las diferentes organizaciones participan por igual en tales proyectos.

Dicha cuestión es analizada en esta pregunta, analizando el número de desarrolladores por proyecto y organización (Fig. 9).

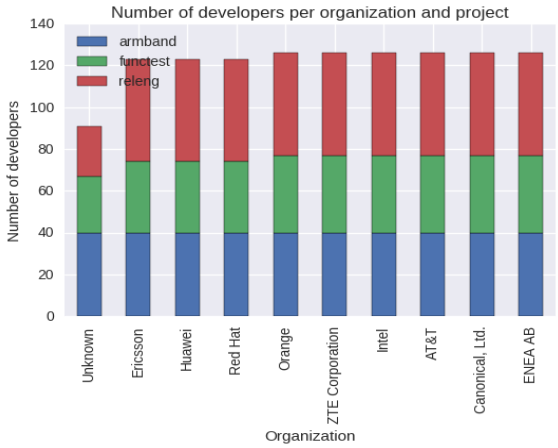


Fig. 9: Número de desarrolladores por organización y proyecto.

Código python query:

s = Search(using=client, index=INDEX).filter('range', author_date={'gt': datetime(2015, 1, 1)})
s.aggs.bucket('by_company', 'terms', field='Commit_org_name').bucket('by_repo', 'terms', field='repo_name', size = 50).metric('Authors', 'cardinality', field='author_name')

Como se puede observar en la anterior figura, los proyectos armgand, functest y releng son de gran importancia para los desarrolladores de las principales organizaciones estudiadas. De manera adicional, se muestra en la siguiente figura el número de autores para los diferentes proyectos en los cuáles participa Ericsson.

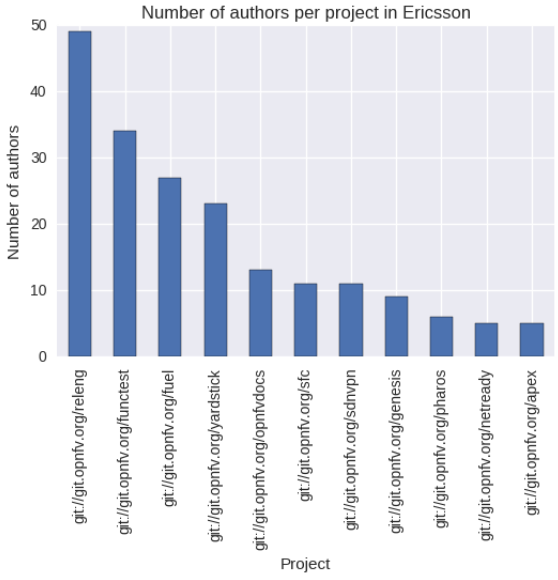


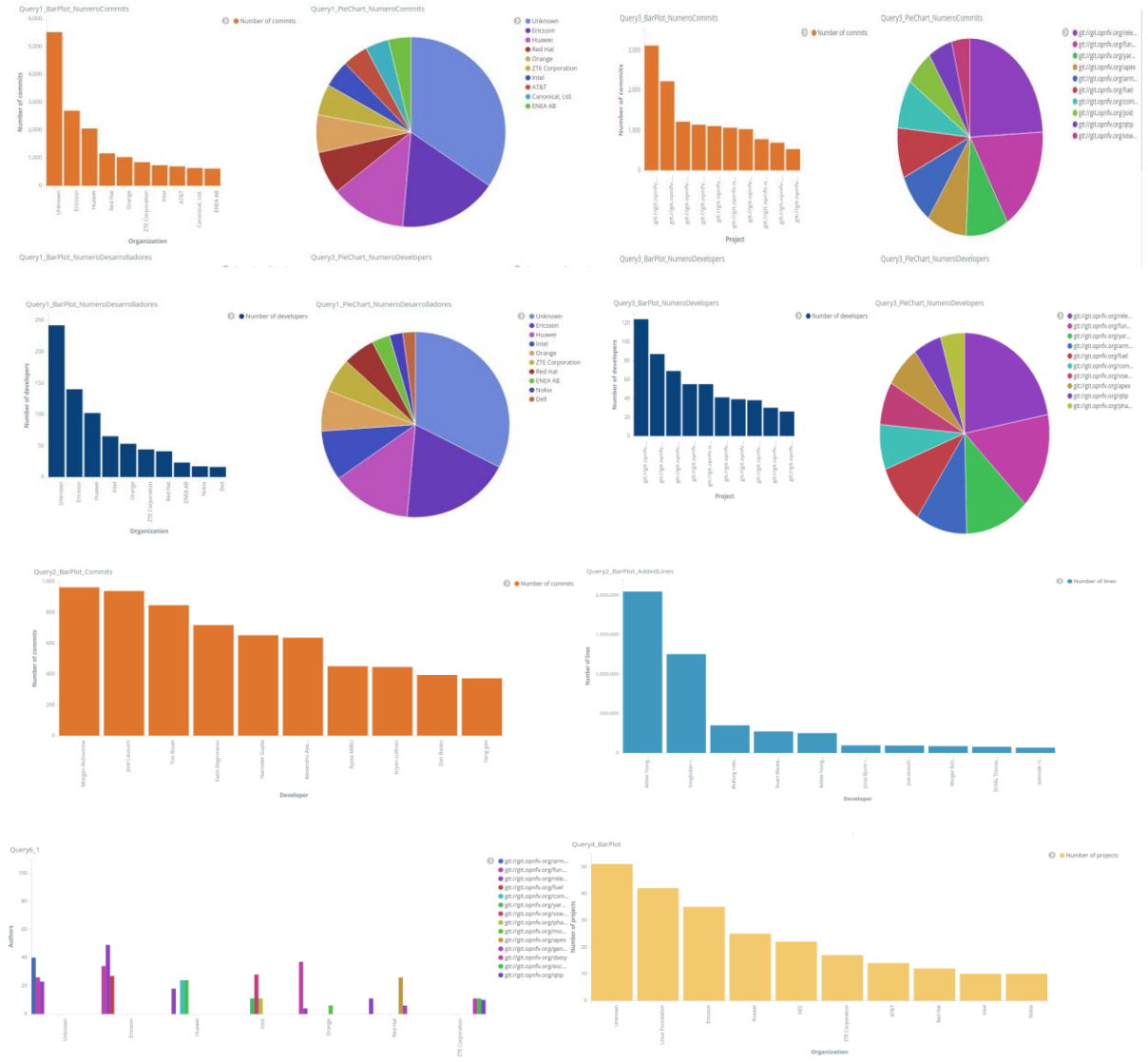
Fig. 10: Número de desarrolladores por proyecto de la empresa Ericsson.

Código python query:

```
company = 'Ericsson'
s = Search(using=client, index=INDEX).filter('range', author_date={'gt': datetime(2015, 1, 1)}).filter("term",Commit_org_name=company)
s.aggs.bucket('by_repo', 'terms', field='repo_name', size=50).metric('Authors', 'cardinality', field='author_name')
```

4.2 Dashboard en Kibana

Todas las anteriores consultas, también han sido resueltas en Kibana. En concreto, se ha desarrollado un dashboard que el lector puede encontrar en la carpeta *visualizaciones_kibana/dashboard*, en el fichero *dashboard.json*. De manera adicional, en la siguiente figura, se incluye una simulación del dashboard obtenido en Kibana.



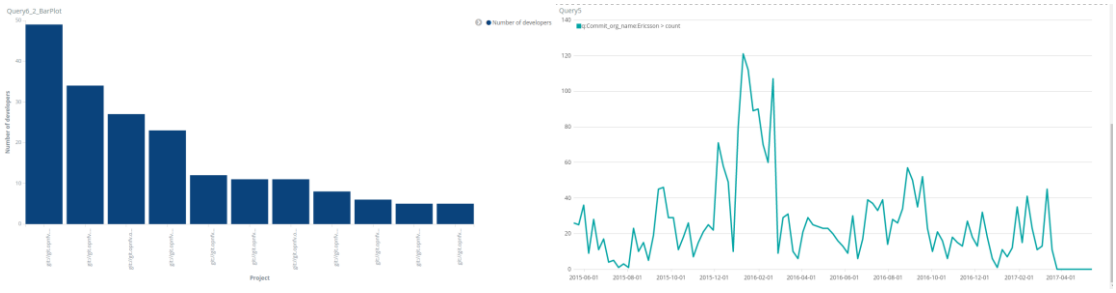


Fig. 11: Simulación del Dashboard de Kibana.

5 Conclusiones

A partir de los resultados anteriores, se pueden extraer las siguientes conclusiones:

- Las organizaciones más importantes son Ericsson y Huawai.
- Los principales desarrolladores son Morgan Richomme y Jose Lausuch.
- Los principales repositorios son releng y functest.
- La organización Linux Foundation presenta un gran número de proyectos, que sin embargo no aúnan demasiada actividad.
- Los desarrolladores de la empresa Ericsson registraron un pico de actividad a finales del año 2015 y principios del año 2016.
- los proyectos armgand, functest y releng son de gran importancia para los desarrolladores de las principales organizaciones estudiadas

6 References

[1] “Open Platform for NVF,” [Online]. Available: https://www.opnfv.org/.