

Análisis de ofertas de empleo mediante datos obtenidos de Twitter e información del INEI



INTEGRANTES:

- MIJAIL DAVIS HUAMAN ROMERO
- EDUARDO HUARCAYA QUISPE
- VICTOR ANDRES CORDOVA BERNUY

Caso de Uso

El presente trabajo tiene por finalidad analizar la relación existente entre las ofertas de empleo publicadas en la red social Twitter versus el tipo de carreras universitarias y técnicas estudiadas por los limeños.

El estudio tendrá como base 04 pilares:

- 1)** Procesamiento de información a través de tecnologías Big Data
- 2)** Recolección masiva de información de la web a través del uso de algoritmos de “Web Scraping”
- 3)** Procesamiento de información a través de algoritmos de aprendizaje autónomo “Machine Learning”
- 4)** Visualización de resultados a través de herramientas como Power BI y Excel.

1. Necesidad de Tecnologías Big Data



El presente trabajo sustentará su análisis con información proveniente de dos grandes fuentes de datos.

- ❖ **Twitter**, considerada una de las mas importantes redes sociales de la actualidad con 330 millones usuarios activos y 500 millones de Tweets por día (*).
- ❖ **INEI**, Instituto Nacional de Estadísticas e informáticas, organismo encargado de los censos de población, empresas, viviendas, etc., en Perú.

Dada la magnitud de información que es posible recolectar se nos hace viable la aplicación de Tecnologías Big Data.

(*) Fuente: <https://www.flimper.com/blog/es/estadisticas-globales-de-twitter-2018->

Análisis de los Datos (3V's)

- ❑ **Velocidad:** Al tener como fuente de información datos proveniente de Twitter, estos se generan constantemente a través de Tweets publicados.
- ❑ **Variedad:** Se procesara información proveniente del INEI (data estructurada) e información proveniente de Twitter (data no estructurada).
- ❑ **Volumen:** INEI y Twitter, gestionan grandes volúmenes de información.

2. Recolección de Información de Twitter



Screenshot of a Twitter profile page for 'Portal Empleos Peru' (@portalemploes). The profile picture shows a man in a suit. The bio reads: 'Empleos profesionales en el Perú. Bolsa de Trabajo. Ingeniería, Arquitectura, Administración, Economía, Contabilidad.' The account has 42,2 mil tweets, 46 following, and 6.669 followers. A yellow circular icon with a person icon and the word 'EMPLEO' is overlaid on the profile picture. The timeline shows two tweets from the user:

Tweets **Tweets y respuestas**

Portal Empleos Peru @portalemploes - 14 min
#Empleo #Peru Practicante Pre profesional Centro Computo UNIBANCA Lima

CV Trabajos: Practicante Pre profesional Centro Co...
Practicante Pre profesional Centro Computo UNIBANCA Lima : <https://www.cvtrabajos.com/2019/07/practicante-pre-profesional-centro.html>
cvtrabajos.com

Portal Empleos Peru @portalemploes - 14 min
#Empleo #Peru Analista TI Jr. SCOTIABANK PERU SAA Lima

¿Nuevo en Twitter?
Regístrate ahora para obtener tu propia cronología personalizada!

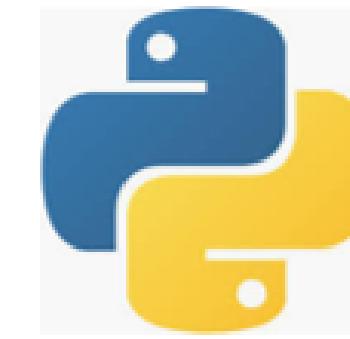
Regístrate

También te puede gustar
Actualizar

Empleos en el Peru
@BuscaEmpleoPeru

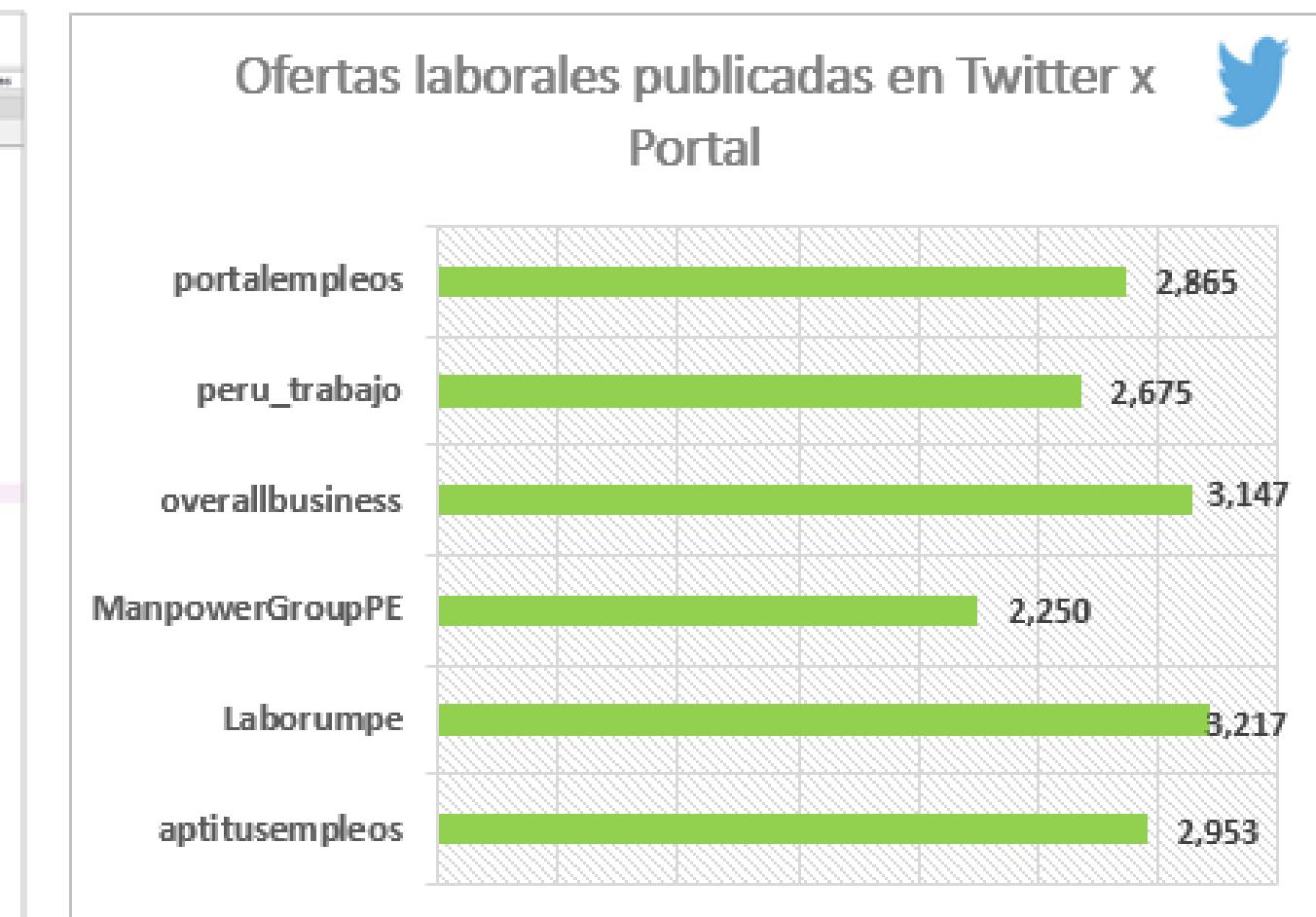
APITUS.com

Web Scraping



Se procedió a recolectar Tweets a través del uso de algoritmos de “Web Scraping” desarrollados en el lenguaje de programación Python.

```
Spyder (Python 3.7)
Archivo Editar Buscar Código fuente Ejecutar Depurar Terminales Proyectos Herramientas Ver Ayuda
Editor -> D:\Documentos\Descargas\Get_tweets.py
temp.py Get_tweets.py
1 import tweepy
2 https://github.com/tweepy/tweepy
3 import os
4 import xlswriter
5 import datetime
6
7 #Twitter API credentials
8 consumer_key = "Yodd7eKtC1zGLwVjFogTulgP"
9 consumer_secret = "JWvVtHfAPsud4TvhU1Dg1uSg8dLhpvLekpvW2P9vJhJ"
10 access_key = "325740640-Uf5uB4qyvhuM0CjRc05r04u4098v1N8Cqf"
11 access_secret = "mpuTH5hgloL0TmPHKochkttCumerDq9Gd4kC5OxIcoB"
12
13 def get_all_tweets(screen_name):
14     #Twitter only allows access to a users most recent 3240 tweets with this method
15
16     auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
17     auth.set_access_token(access_key, access_secret)
18     api = tweepy.API(auth)
19
20     #Initialize a list to hold all the tweets
21     alltweets = []
22     new_tweets = []
23     outtweets = []
24
25     #make initial request for most recent tweets (200 is the maximum allowed count)
26     new_tweets = api.user_timeline(screen_name = screen_name,count=200)
27
28     #Save most recent tweets
29     alltweets.extend(new_tweets)
30
31     #Save the id of the oldest tweet less one
32     oldest = alltweets[-1].id - 1
33
34     #Keep grabbing tweets until there are no tweets left to grab
35     while len(new_tweets) > 0:
36         print("getting tweets before %s" % (oldest))
37
38         #all subsequent requests use the max_id param to prevent duplicates
39         new_tweets = api.user_timeline(screen_name = screen_name,count=200,max_id=oldest)
```

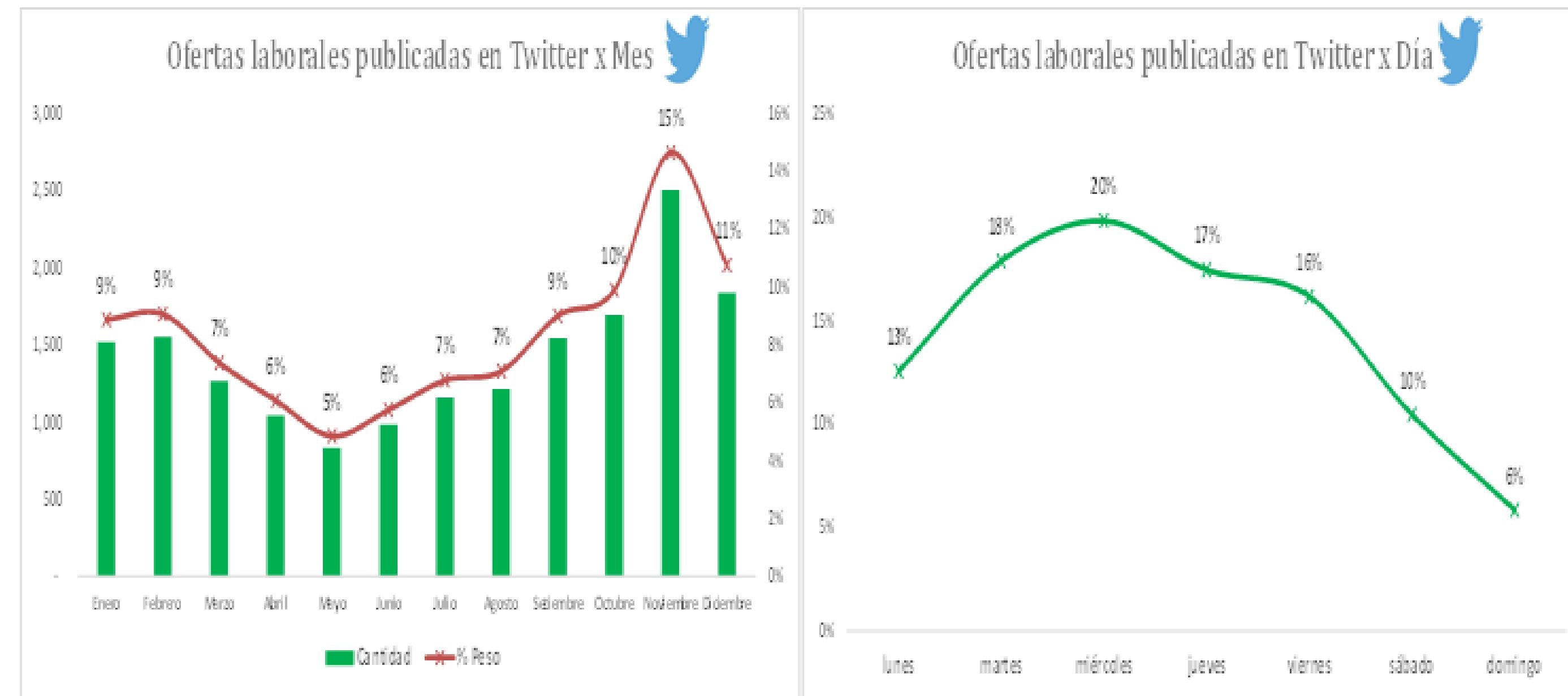


Fuente: Tweets de portales de empleo en Perú 2012-2018

Tweets Descargados



Tweets Descargados



ENAHO 2018

 INSTITUTO NACIONAL DE ESTADÍSTICA E INFORMATICA

MICRODATOS

BASE DE DATOS

Inicio Consulta por Encuestas Documentación

[PRESENTACIÓN](#) [GUÍA DE USUARIO](#)

CONSULTA POR ENCUESTA

Sinrase seleccionar Encuesta, Año y Periodo y a continuación se mostrarán todos los Módulos de la Encuesta Seleccionada. Luego proceda a descargar el módulo de su interés.

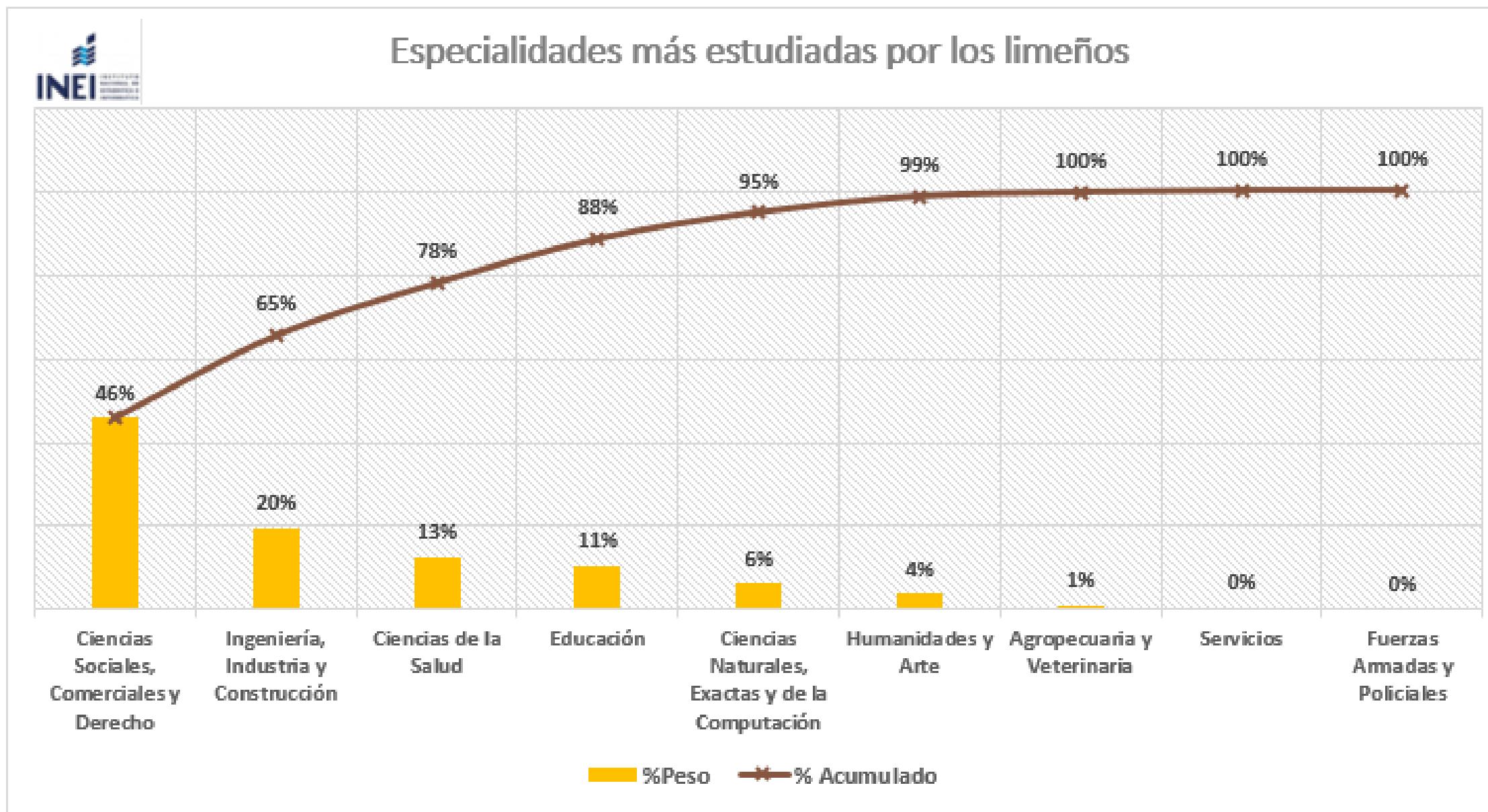
ENCUESTA ENAHO Metodología ACTUALIZADA

Condiciones de Vida y Pobreza - ENAHO

AÑO: 2018 Periodo: Anual - (Ene-Dic)

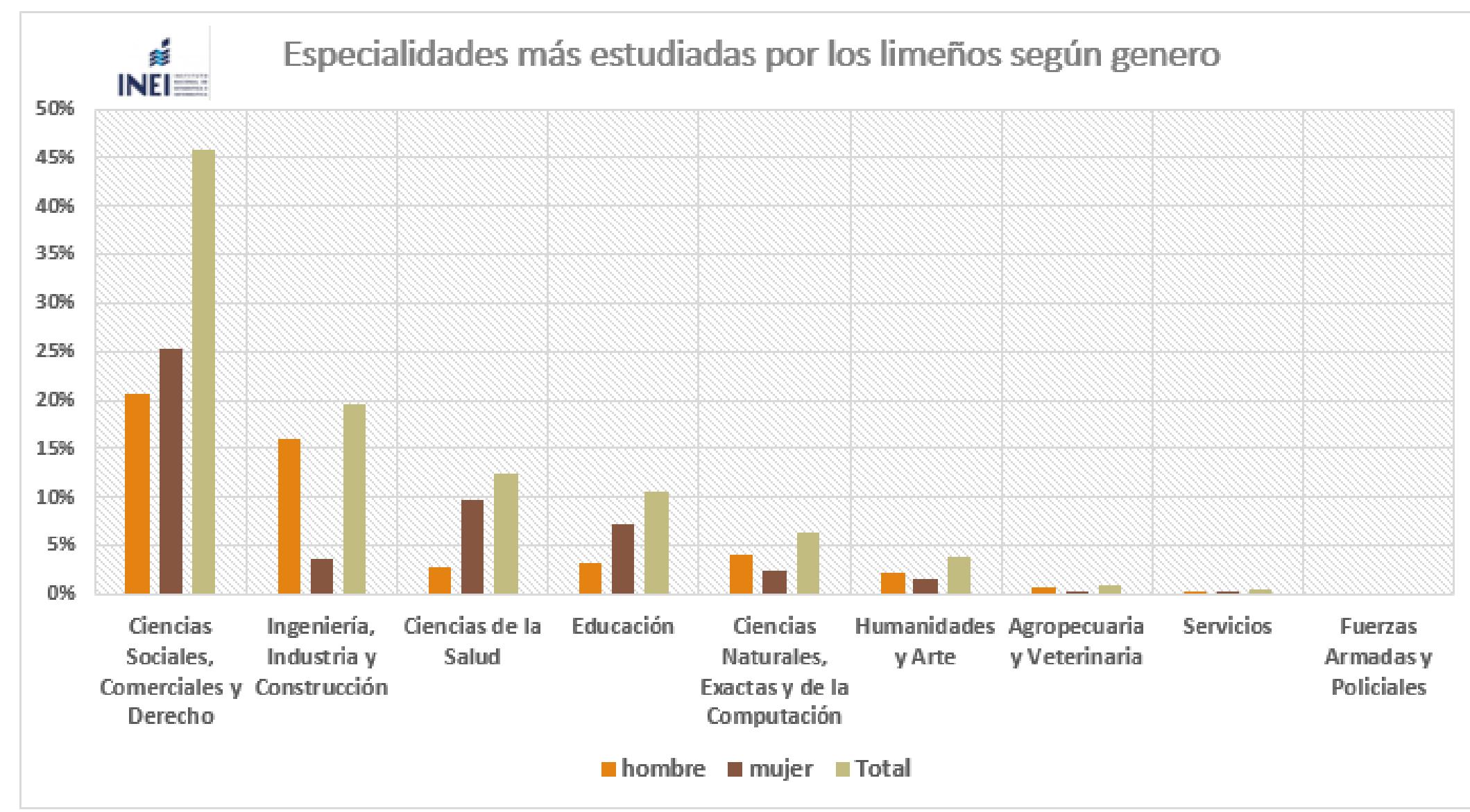
Nro	Año	Periodo	Código Encuesta	Encuesta	Código Modulo	Módulo	Ficha	Descarga
1	2018	65	634	Condiciones de Vida y Pobreza - ENAHO	1	Características de la Vivienda y del Hogar		
2	2018	65	634	Condiciones de Vida y Pobreza - ENAHO	2	Características de los Miembros del Hogar		
3	2018	65	634	Condiciones de Vida y Pobreza - ENAHO	3	Educación		
4	2018	65	634	Condiciones de Vida y Pobreza - ENAHO	4	Salud		
5	2018	65	634	Condiciones de Vida y Pobreza - ENAHO	5	Empleo e Ingresos		
6	2018	65	634	Condiciones de Vida y Pobreza - ENAHO	7	Gastos en Alimentos y Bebidas (Módulo 601)		
7	2018	65	634	Condiciones de Vida y Pobreza - ENAHO	8	Instituciones Beneficas		
8	2018	65	634	Condiciones de Vida y Pobreza - ENAHO	9	Mantenimiento de la Vivienda		
9	2018	65	634	Condiciones de Vida y Pobreza - ENAHO	10	Transportes y Comunicaciones		
10	2018	65	634	Condiciones de Vida y Pobreza - ENAHO	11	Servicios a la Vivienda		

Especialidades más estudiadas por los limeños



Fuente: INEI

Especialidades más estudiadas por los limeños según género



Fuente: INEI

3. Procesamiento de información a través de Machine Learning



Algoritmo TF – IDF

Algoritmo que nos ayuda a medir la relevancia de las palabras en una colección de documentos (Tweets).

$$\text{tfidf}_{i,j} = \text{tf}_{i,j} \times \log \left(\frac{N}{\text{df}_i} \right)$$

$\text{tf}_{i,j}$ = total number of occurrences of i in j
 df_i = total number of documents (speeches) containing i
N = total number of documents (speeches)

```
In [32]: # Running LDA using TF-IDF
lda_model_tfidf = gensim.models.LdaMulticore(corpus_tfidf, num_topics=10, id2word=dictionary, passes=2, workers=4)
for idx, topic in lda_model_tfidf.print_topics(-1):
    print('Topic: {} Word: {}'.format(idx, topic))

# Nuevamente, puedes distinguir los diferentes tópicos usando las palabras en cada tópico y sus pesos correspondientes?
```

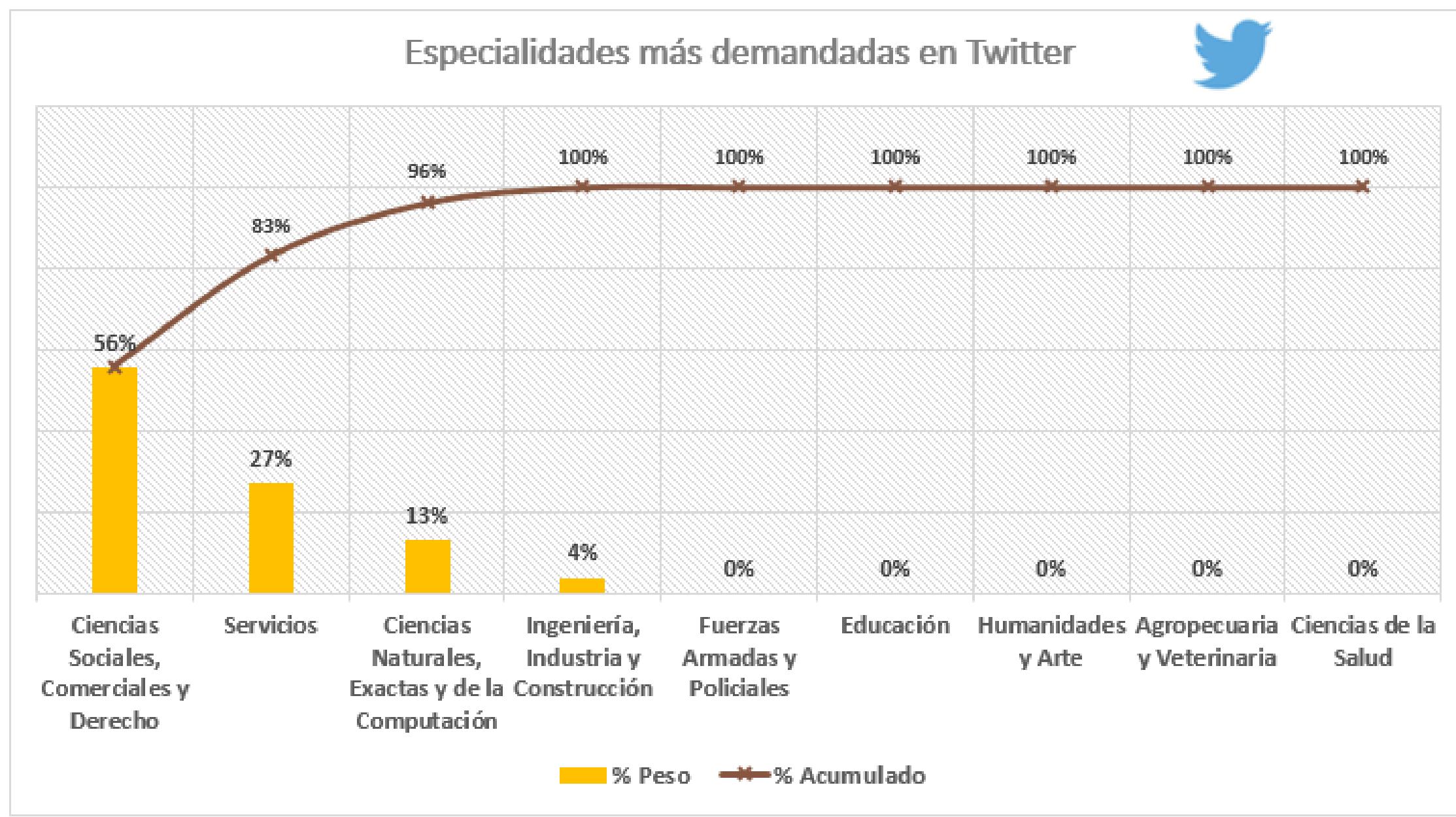
Análisis de resultados

```
In [32]: # Running LDA using TF-IDF
lda_model_tfidf = gensim.models.LdaMulticore(corpus_tfidf, num_topics=10, id2word=dictionary, passes=2, workers=4)
for idx, topic in lda_model_tfidf.print_topics(-1):
    print('Topic: {} Word: {}'.format(idx, topic))

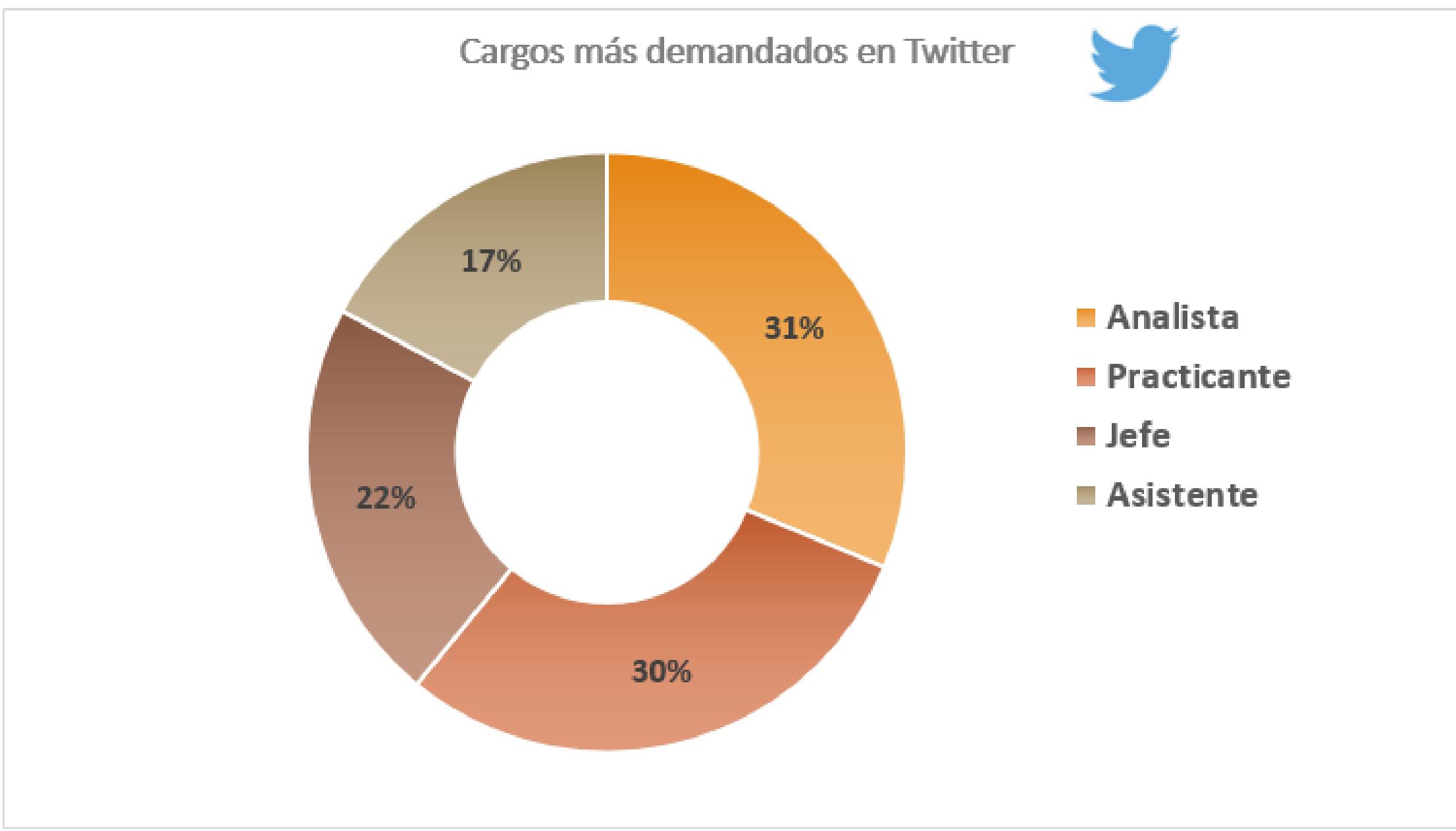
# Nuevamente, puedes distinguir los diferentes tópicos usando las palabras en cada tópico y sus pesos correspondientes?
```

Topic: 0 Word: 0.020*"perú" + 0.019*"portaltrabajo" + 0.015*"empleo" + 0.014*"transna" + 0.013*"para" + 0.013*"empresa" + 0.010*"lima" + 0.009*"búsqueda" + 0.009*"postula" + 0.008*"limpieza"
Topic: 1 Word: 0.018*"overal" + 0.016*"para" + 0.016*"corporativo" + 0.016*"oportunidad" + 0.013*"comienza" + 0.013*"promocion" + 0.013*"área" + 0.013*"trabajo" + 0.012*"portaltrabajo" + 0.012*"administrativa"
Topic: 2 Word: 0.020*"perú" + 0.018*"portaltrabajo" + 0.018*"asesor" + 0.016*"comerci" + 0.015*"experiencia" + 0.015*"empleo" + 0.014*"transna" + 0.013*"empresa" + 0.011*"teleoperador" + 0.011*"impulsamo"
Topic: 3 Word: 0.014*"humano" + 0.012*"facebook" + 0.011*"para" + 0.011*"nuestra" + 0.010*"analista" + 0.010*"overal" + 0.009*"trabajo" + 0.009*"recurso" + 0.008*"profesion" + 0.007*"oscar"
Topic: 4 Word: 0.028*"portal" + 0.026*"completo" + 0.025*"perfil" + 0.023*"postular" + 0.021*"pued" + 0.017*"vshyfm" + 0.015*"overal" + 0.015*"ingresando" + 0.014*"visita" + 0.013*"empleo"
Topic: 5 Word: 0.022*"week" + 0.016*"overal" + 0.015*"éxito" + 0.012*"follow" + 0.012*"familia" + 0.011*"asistent" + 0.011*"perú" + 0.011*"practicant" + 0.010*"lleva" + 0.009*"empresa"
Topic: 6 Word: 0.023*"perú" + 0.017*"empleo" + 0.016*"para" + 0.015*"empresa" + 0.014*"transna" + 0.010*"overal" + 0.009*"trabajo" + 0.008*"reclutamiento" + 0.008*"industri" + 0.008*"tien"
Topic: 7 Word: 0.019*"perú" + 0.018*"twitter" + 0.013*"overal" + 0.012*"empleo" + 0.011*"desarrollo" + 0.011*"transna" + 0.011*"empresa" + 0.010*"impulsamo" + 0.010*"cajero" + 0.009*"correo"
Topic: 8 Word: 0.063*"perú" + 0.040*"transna" + 0.036*"empleo" + 0.035*"empresa" + 0.020*"lima" + 0.016*"semana" + 0.015*"venta" + 0.015*"buen" + 0.011*"vendedor" + 0.010*"inicio"
Topic: 9 Word: 0.016*"perú" + 0.016*"empresa" + 0.016*"client" + 0.016*"programador" + 0.015*"nuestro" + 0.014*"jefe" + 0.012*"empleo" + 0.012*"venta" + 0.011*"servicio" + 0.010*"analista"

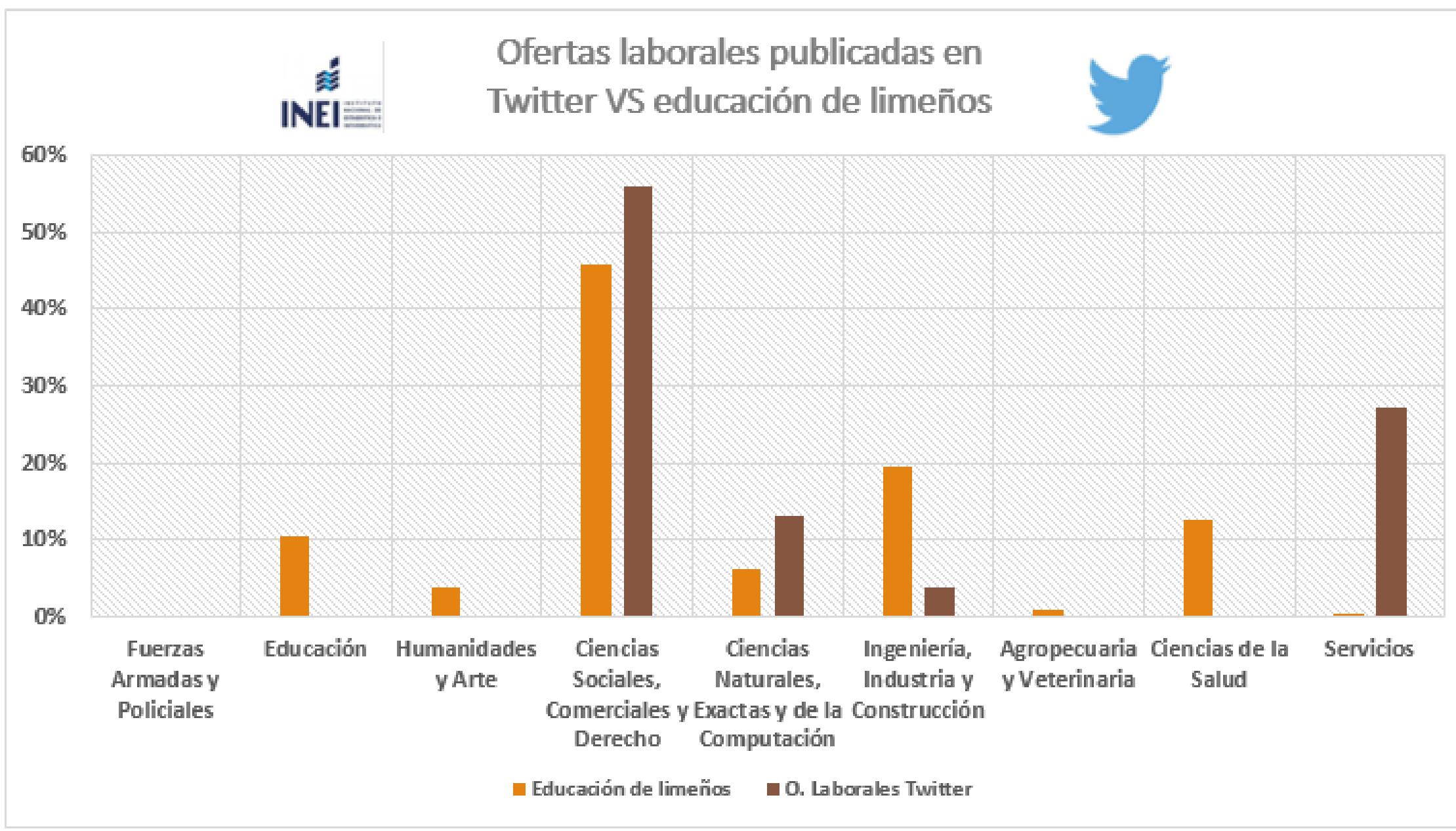
Especialidades mas demandadas en Twitter



Cargos mas demandados en Twitter



4. Visualización de Resultados



Anexos:

- ❑ Tablero: Actividades preliminares
- ❑ Tablero: Actividades de desarrollo

Tablero: Actividades preliminares (Tablero Scrum)

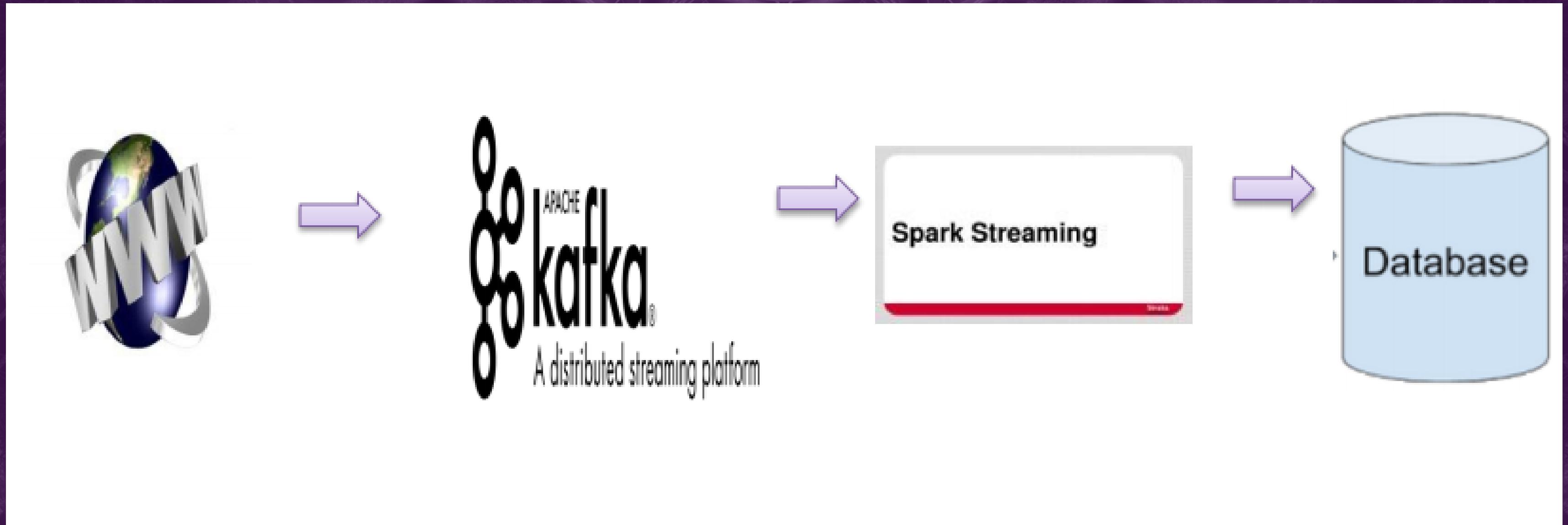
#	Actividades preliminares	Responsable	To do	Doing	Done
1	Enviar lista de cuentas twitter para la búsqueda de empleos en Perú	Víctor			x
2	Enviar lista de portales de trabajo para la búsqueda de empleo en Perú	Víctor			x
3	Enviar lista de cuentas twitter de los principales portales de noticias en Perú	Víctor			x
4	Enviar lista de cuentas twitter de las principales universidades e institutos del país, clasificar instituciones por el departamento donde operan (agregar columnas con tipo de institución (universidad o instituto) y ubicación)	Víctor			x
5	Enviar lista de cuentas twitter de las principales empresas del país, clasificar empresas por sectores económicos y principal departamento en donde opera (agregar columnas con sectores económicos e ubicación)	Víctor			x
6	Investigar sobre el uso de algoritmos web scraping, traer páginas y ejemplos específicos encontrados en la web	Víctor			x
7	Investigar sobre el uso de Power BI utilizando como fuente de datos Hadoop (tablasHive, etc.)	Víctor			x
8	Enviar lista de atributos de información a encontrar en la ENAHO 2018	Mijail			x
9	Descargar información ENAHO 2018 y guardarla en formato csv, parquet u otro formato que pueda subirse a Google Platform y explotarse con Hadoop (Hive, Spark, etc.)	Mijail			x
10	Construir primera versión del caso de uso, explicar métricas e indicadores a construirse, ademas de justificar la aplicación del Big Data (3Vs)	Eduardo			x

Tablero: Actividades de desarrollo (Tablero Scrum)

#	Actividades de desarrollo	Responsable	To do	Doing	Done
1	Desarrollar la presentación en Jupyter, Jupyter Slides, Slides, Swamp, etc.	Víctor			x
2	Desarrollar el tablero Scrum con todas las actividades que realizarán para desarrollar el proyecto	Eduardo			x
3	Explicar caso de uso	Eduardo			x
4	Exponer un caso de uso de negocio donde se necesite aplicar tecnologías de big data y que tenga al menos un flujo Streaming	Eduardo			x
5	Justificar la necesidades por lo que usaran de Big Data (3vs?)	Eduardo			x
6	Identificar las fuentes que necesitarían (externas o internas)	Mijail			x
7	Sizing de la generación de data por minuto, día, hora y mensual	Víctor			x
8	Clasificar correctamente las fuentes (tradicional, no tradicional, estructurada, semi estructurada, no estructurada, etc.)	Mijail			x
9	Diseñar la arquitectura conceptual de la solución (capas de data lake)	Todos			x
10	Diseñar la arquitectura tecnológica de la solución (Usare hdfs, spark, kafka,...)	Todos			x
11	Justificar los perfiles de Big Data necesarios	Todos			x
12	Ingestar una o mas fuentes a un DataLake (Google cloud) utilizando comandos HDFS (incluir screenshots)	Todos			x
13	Crear al menos 4 tablas en Hive con la data cargada, una simple, con partición dinámica, otra con partición estática y bucketing	Todos			x
14	Realizar transformaciones simples con funciones UDFs nativas Apache Hive	Todos			x
15	Realizar transformaciones complejas (joins, aggregates, etc.) con Apache Spark	Todos			x
16	Generar al menos dos tablas con Apache HBase, no necesariamente tiene que guardar relación de negocio con todo el caso de uso	Todos			x
17	Explicar teóricamente como funcionaría el flujo en streaming usando Apache Kafka	Víctor			x
18	Explicar teóricamente como funcionaría el flujo en streaming usando Apache Flume*	Víctor			x
19	Guardar archivos, código, presentación, informe en la plataforma Github	Víctor			x
20	Analizar la data con Pyspark (Python con Spark)	Mijail			x
21	Mostrar gráficas y/o indicadores en Power BI o Excel	Eduardo			x
22	Publicar articulo en Medium	Eduardo			x

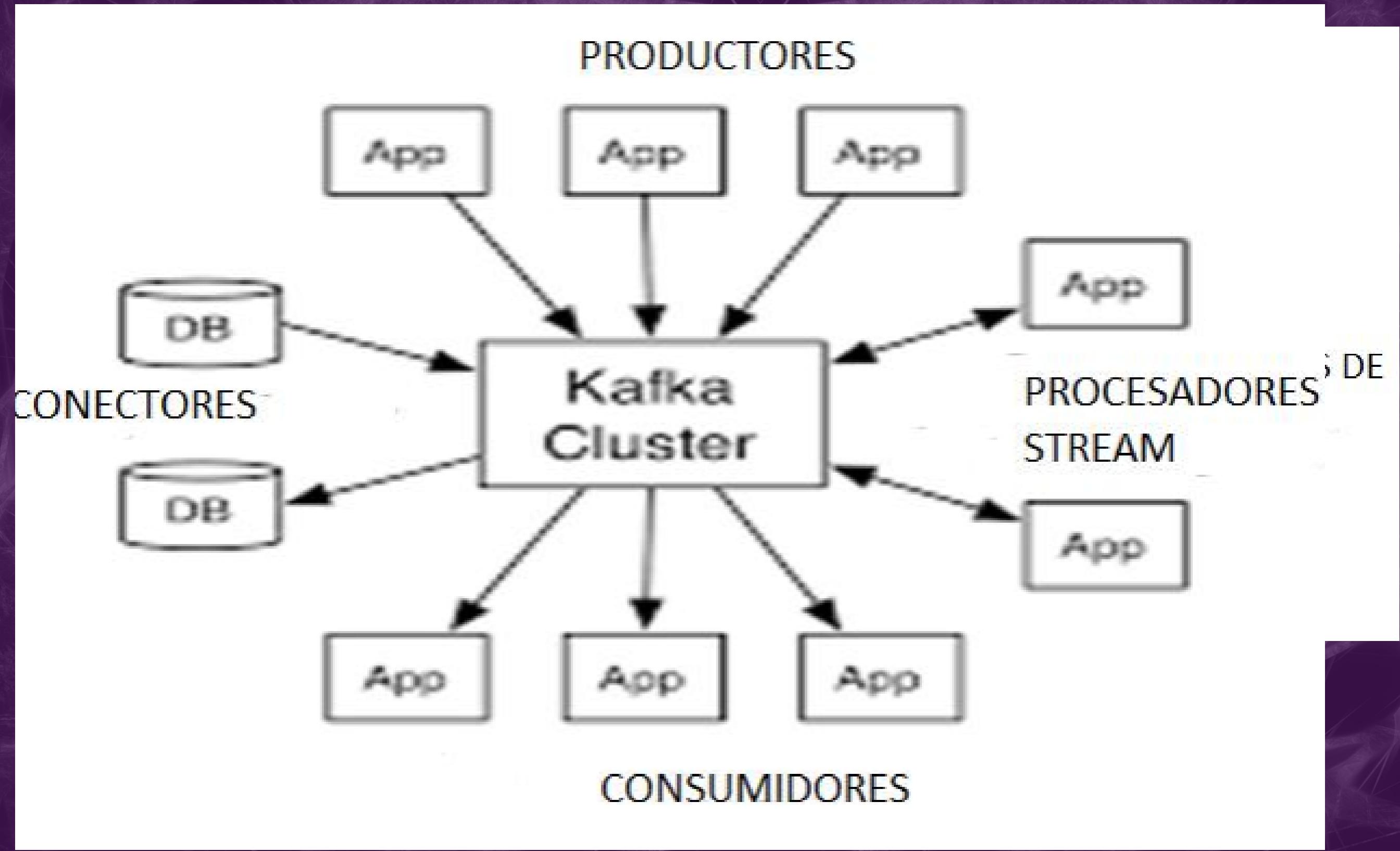
ARQUITECTURA STREAMING CON KAFKA

A nivel de diagrama de Caja Nagra



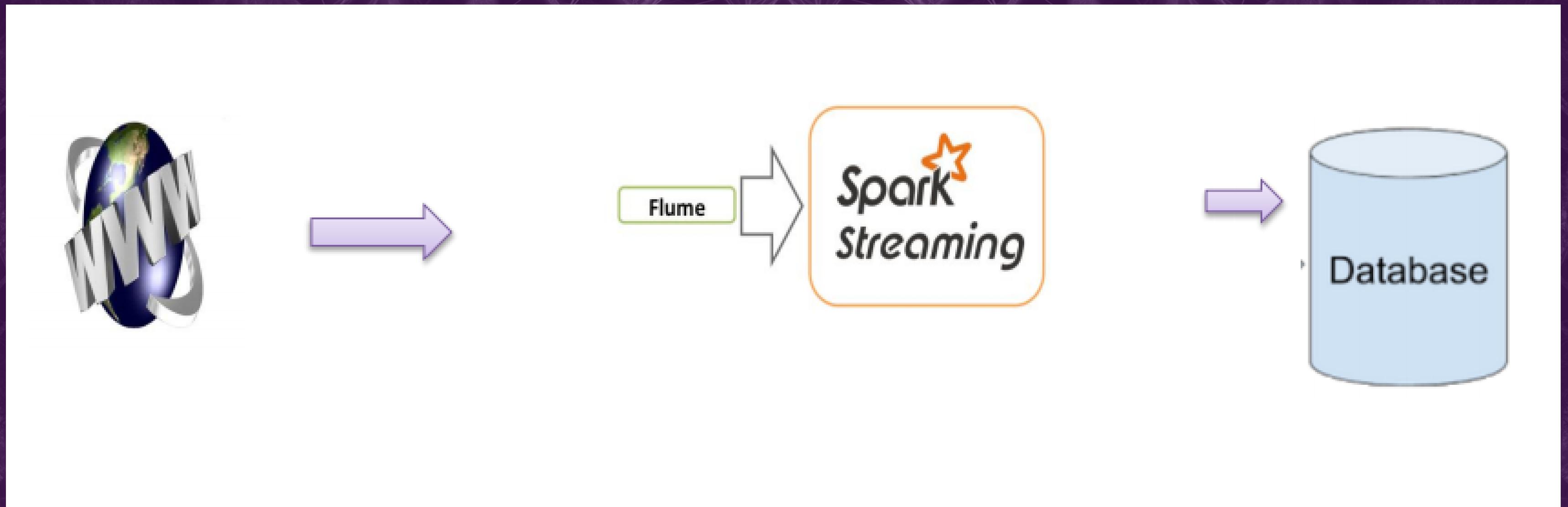
ARQUITECTURA STREAMING CON KAFKA

A nivel de caja blanca



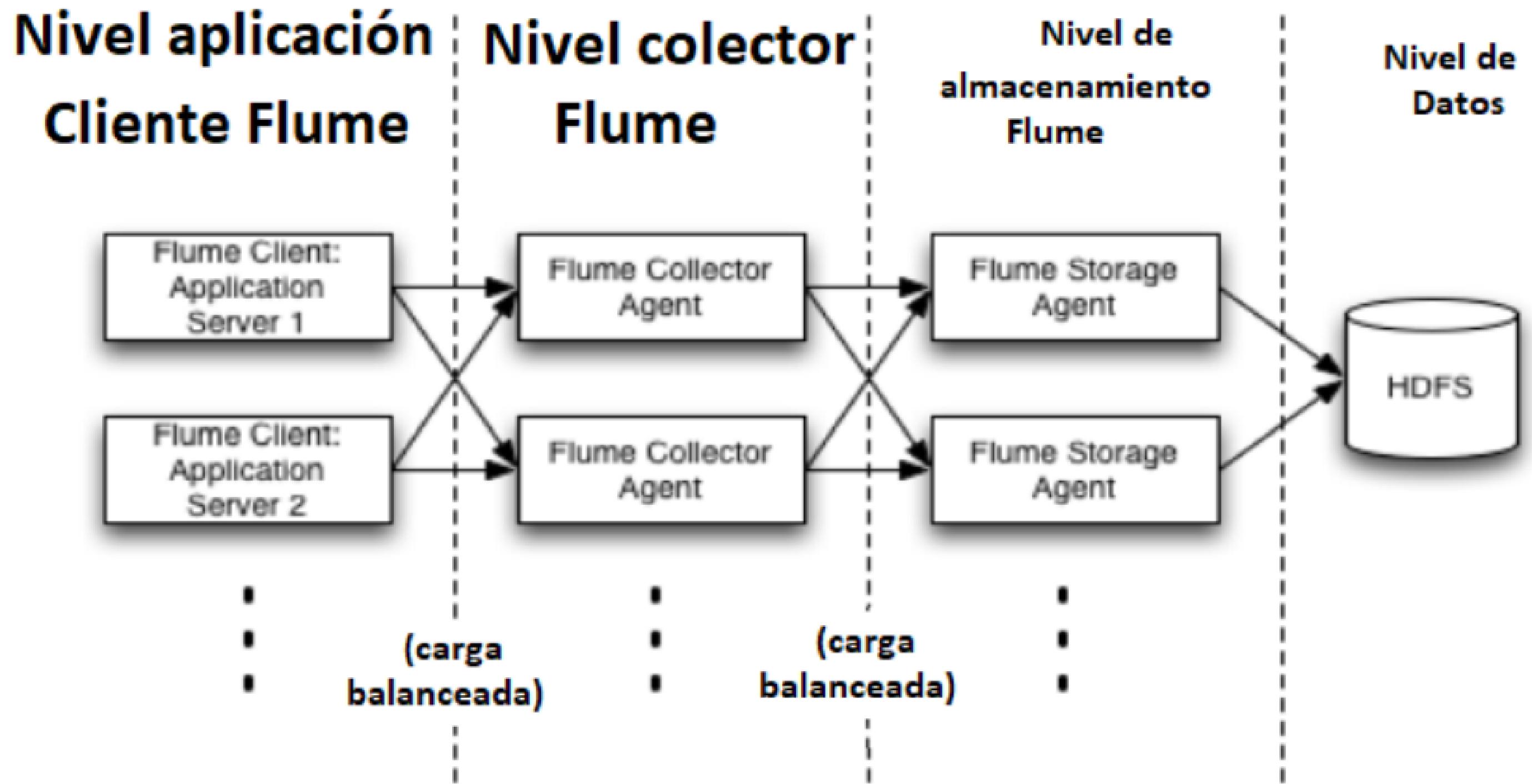
Arquitectura Streaming con Apache Flume

A nivel de caja negra



A nivel de diagrama de caja blanca

Una topología de bloques Flume



GRACIAS TOTALES





MUCHAS GRACIAS
CTIC