

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



Inicio Sobre nosotros

Buscar en Twitter

¿Tienes cuenta? [Iniciar sesión](#)



Tweets 42,2 mil Siguiendo 46 Seguidores 6.669

[Seguir](#)

Portal Empleos Peru
@portalempleos

Empleos profesionales en el Perú. Bolsa de Trabajo, Ingeniería, Arquitectura, Administración, Economía, Contabilidad.

Peru

cvtrabajos.com

Se unió en agosto de 2010

Tweets Tweets y respuestas

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

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

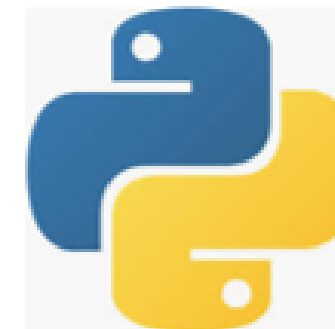
[Regístrate](#)

También te puede gustar.
Actualizar

Empleos en el Peru
@BuscaEmpleoPeru

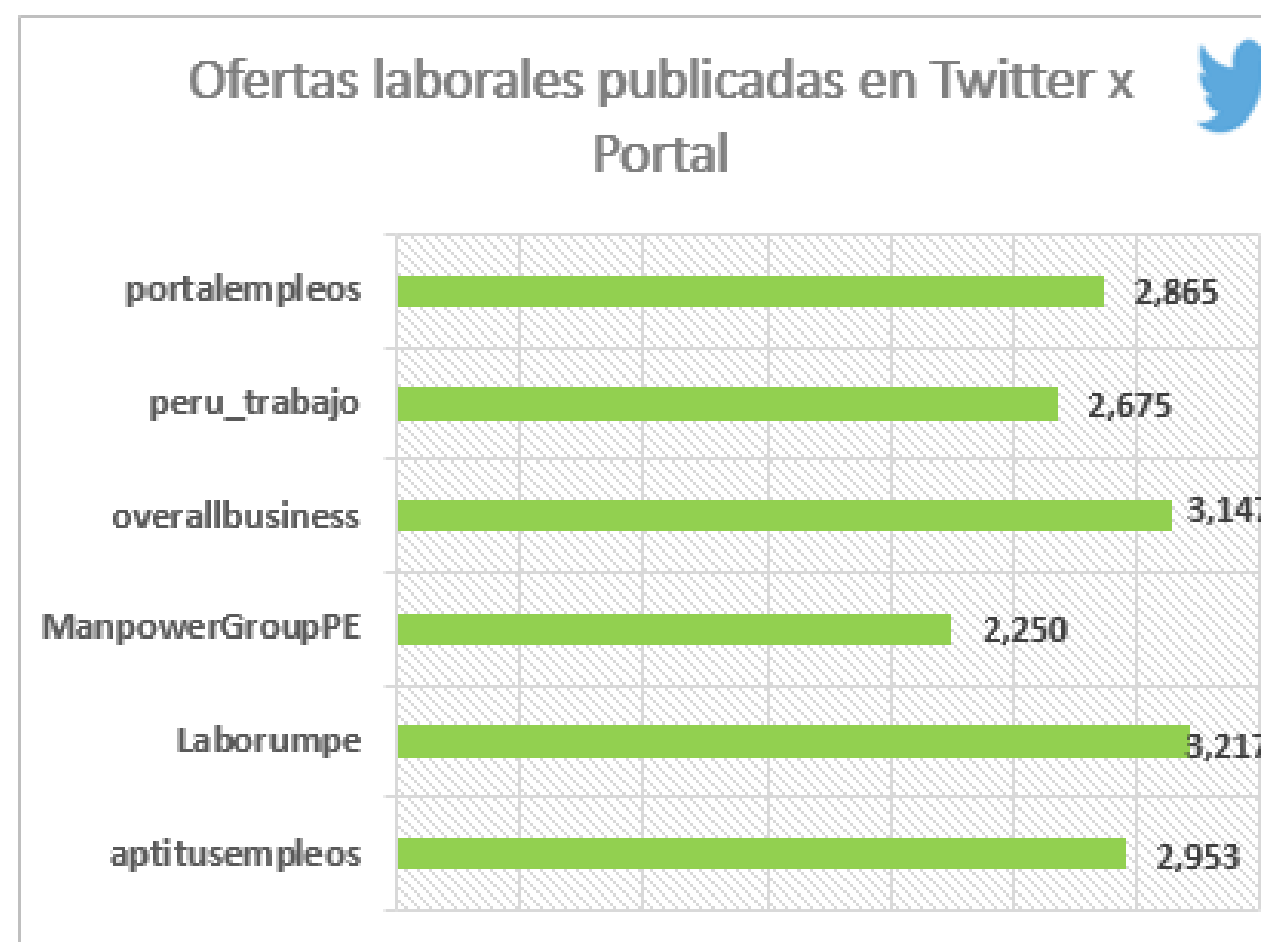
APTITUS.com
@aptituscom

Web Scrapping



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

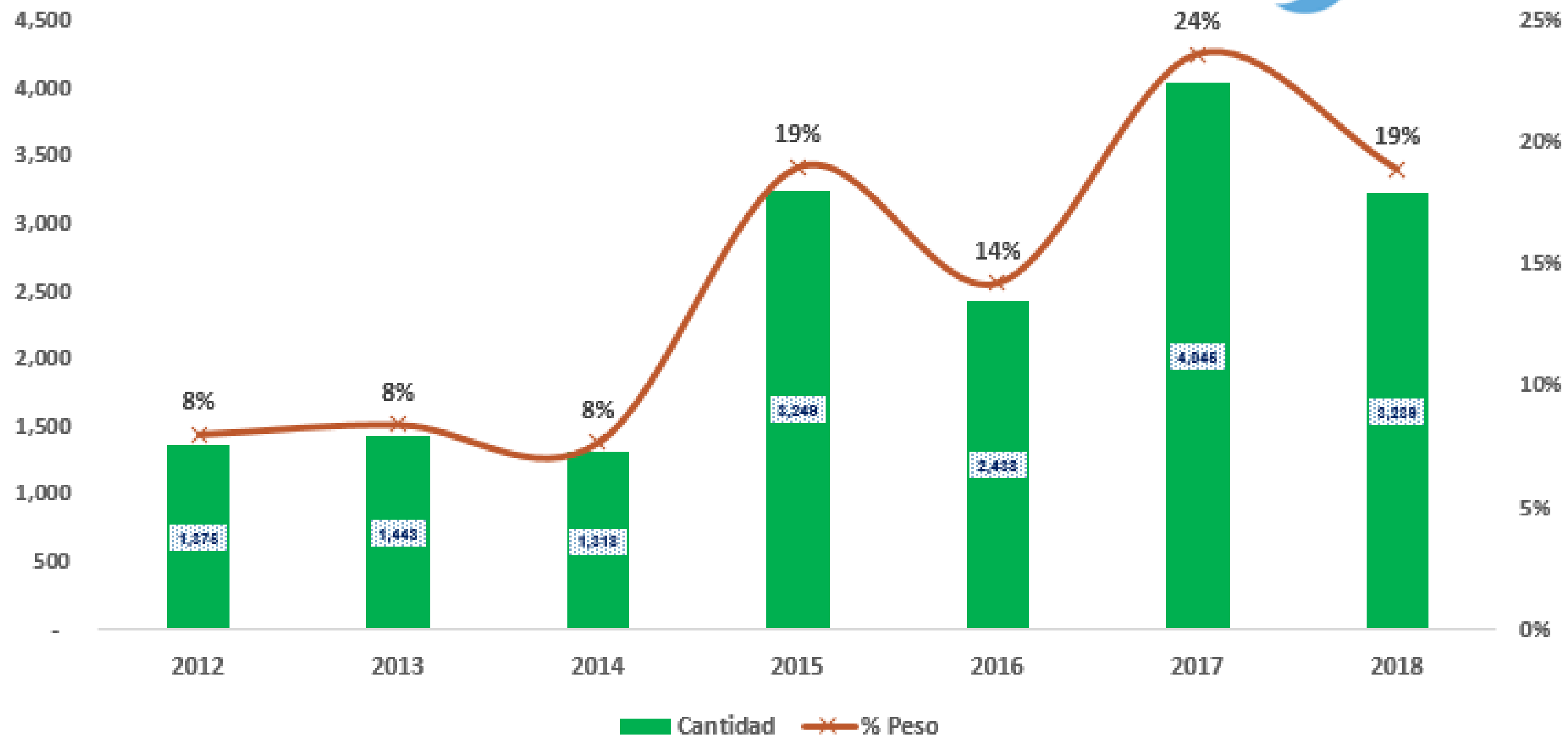
```
Spyder (Python 3.7)
Archivo  Editor  Buscar  Ejecutar  Depurar  Terminales  Proyectos  Herramientas  Ver  Ayuda
Editor - D:\Documentos\Descargas\Get_tweets.py
temp.py  Get_tweets.py
4 import tweepy @https://github.com/tweepy/tweepy
5 import csv
6 import xlswriter
7 import datetime
8
9
10 #Twitter API credentials
11 consumer_key = "9o447eKuCz1nG4wVjfcG3u8gP"
12 consumer_secret = "jYCV9Tf1AP64u4F63VpHJ38g1uF5g841hqvheqzVb3F9v3h3"
13 access_key = "325743640-U7F5u8BmQ7u4h6u8GjRoDS-05uP6d9G8V1M8CqM"
14 access_secret = "xapuTH5hg6oLDTaHFKxhktTCumeDq9W6dk6GSOxllcoB"
15
16
17 def get_all_tweets(screen_name):
18     #Twitter only allows access to a users most recent 3240 tweets with this method
19
20     #authorize twitter, initialize tweepy
21     auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
22     auth.set_access_token(access_key, access_secret)
23     api = tweepy.API(auth)
24
25     #initialize a list to hold all the tweepy Tweets
26     alltweets = []
27     new_tweets = []
28     outtweets = []
29
30     #make initial request for most recent tweets (200 is the maximum allowed count)
31     new_tweets = api.user_timeline(screen_name = screen_name,count=200)
32
33     #save most recent tweets
34     alltweets.extend(new_tweets)
35
36     #save the id of the oldest tweet less one
37     oldest = alltweets[-1].id - 1
38
39     #keep grabbing tweets until there are no tweets left to grab
40     while len(new_tweets) > 0:
41         print("getting tweets before %s" % (oldest))
42
43         #all subsequent requests use the max_id param to prevent duplicates
44         new_tweets = api.user_timeline(screen_name = screen_name,count=200,max_id=oldest)
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
```



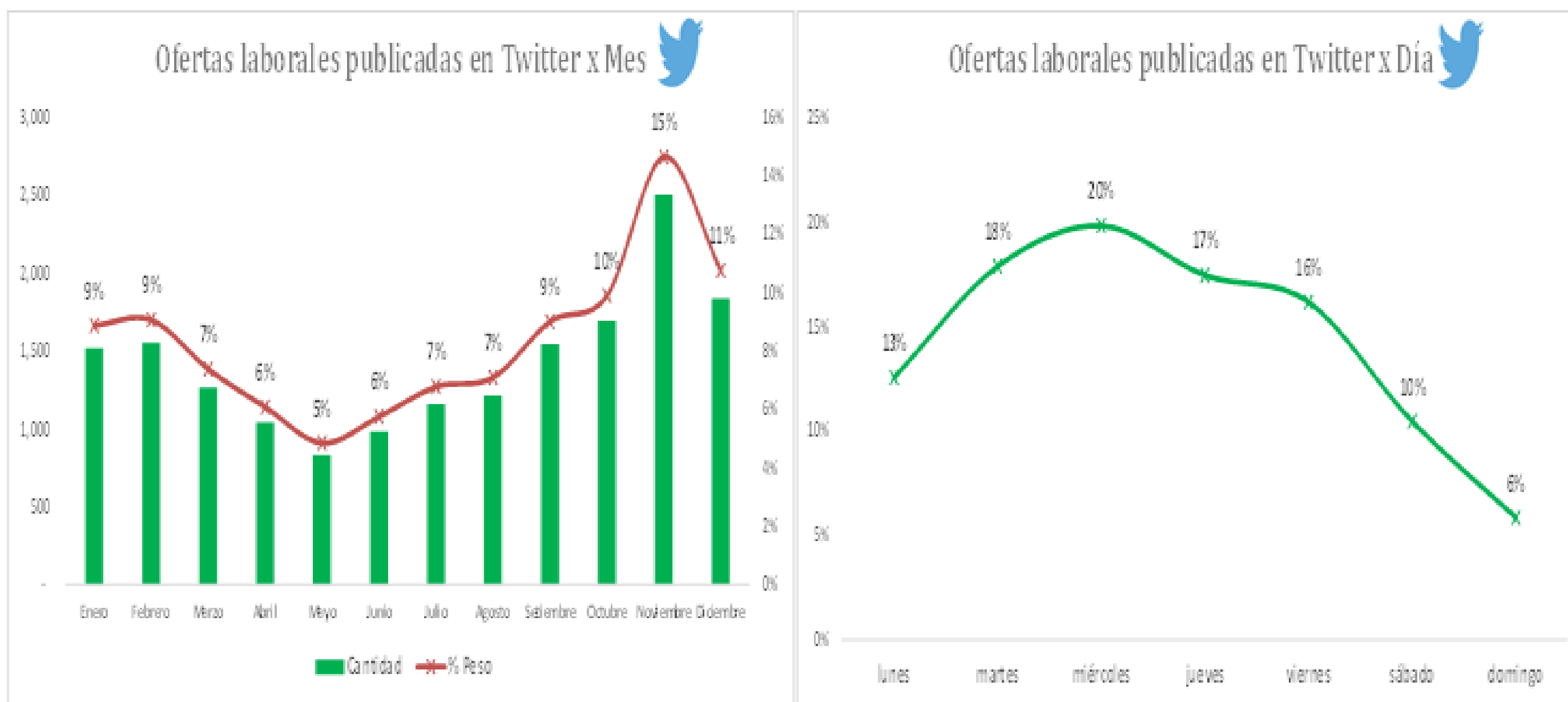
Fuente: Tweets de portales de empleo en Perú 2012-2018

Tweets Descargados

Ofertas laborales publicadas en Twitter x Año



Tweets Descargados



ENAH0 2018



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

CONSULTA POR ENCUESTA

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

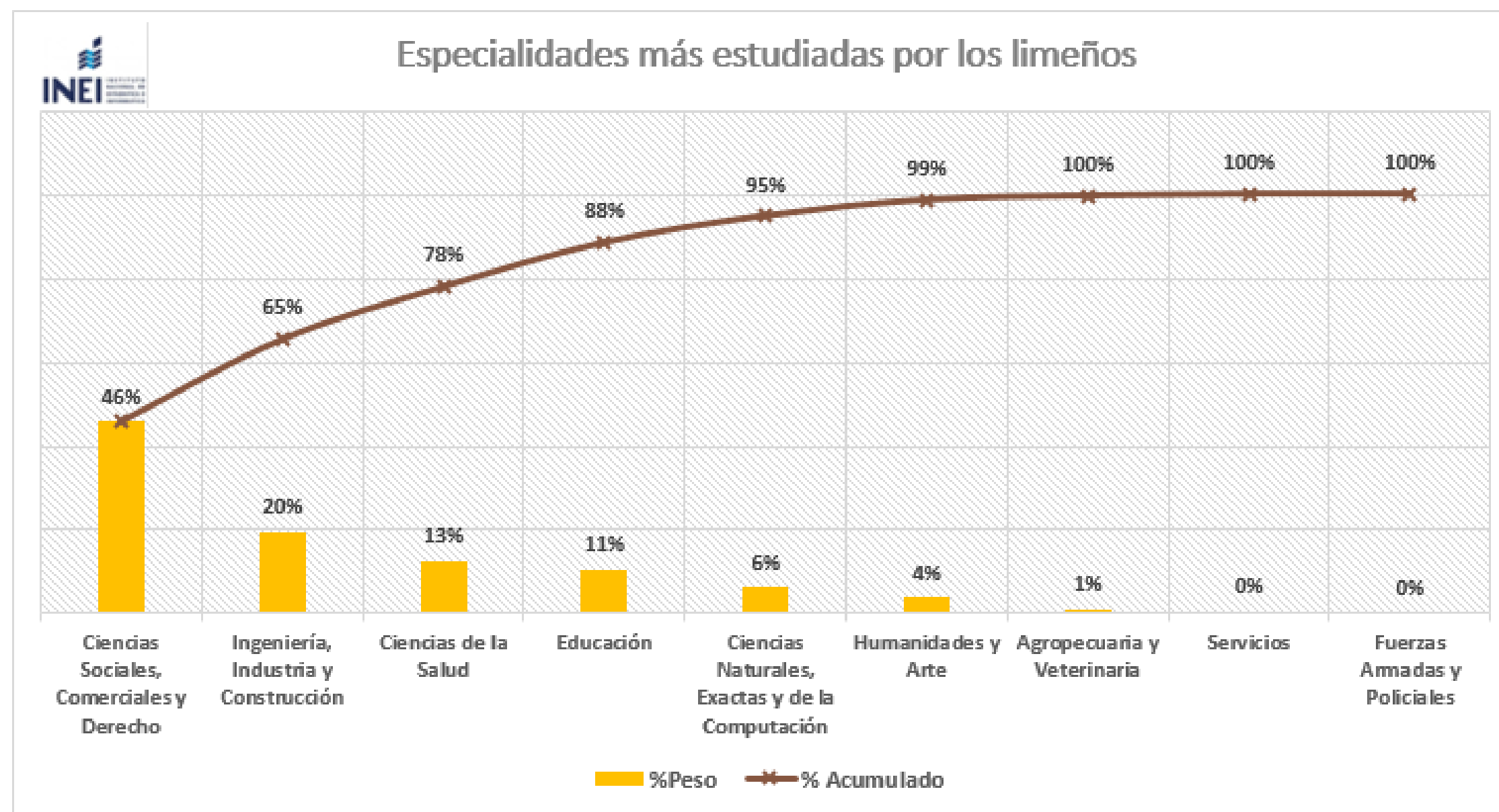
ENCUESTA:

AÑO:

Periodo:

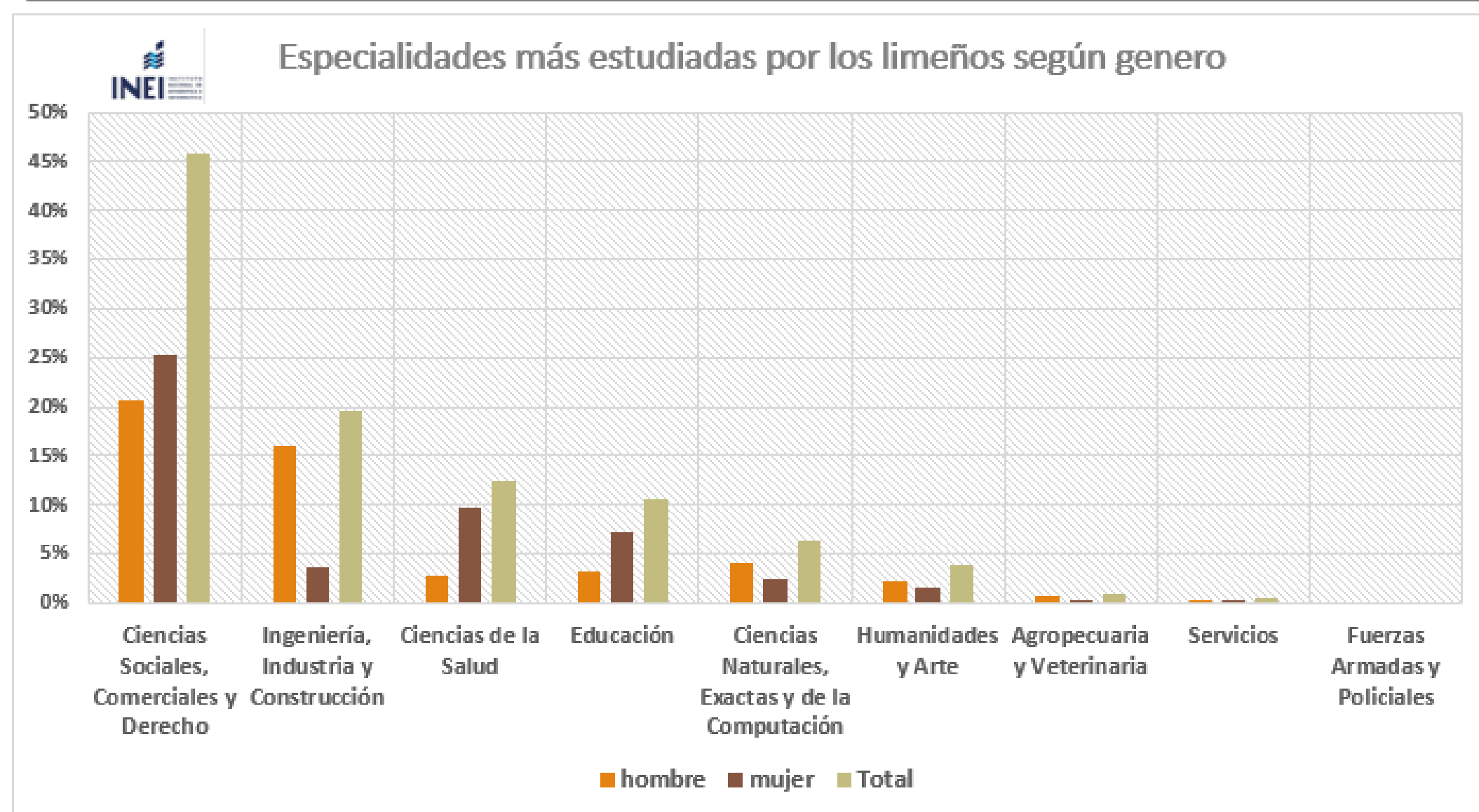
Nro	Año	Período	Código Encuesta	Encuesta	Código Módulo	Módulo	Ficha	Descarga
1	2018	55	634	Condiciones de Vida y Pobreza - ENAH0	1	Características de la Vivienda y del Hogar		
2	2018	55	634	Condiciones de Vida y Pobreza - ENAH0	2	Características de los Miembros del Hogar		
3	2018	55	634	Condiciones de Vida y Pobreza - ENAH0	3	Educación		
4	2018	55	634	Condiciones de Vida y Pobreza - ENAH0	4	Salud		
5	2018	55	634	Condiciones de Vida y Pobreza - ENAH0	5	Empleo e Ingresos		
6	2018	55	634	Condiciones de Vida y Pobreza - ENAH0	7	Gastos en Alimentos y Bebidas (Módulo 601)		
7	2018	55	634	Condiciones de Vida y Pobreza - ENAH0	8	Instituciones Benéficas		
8	2018	55	634	Condiciones de Vida y Pobreza - ENAH0	9	Mantenimiento de la Vivienda		
9	2018	55	634	Condiciones de Vida y Pobreza - ENAH0	10	Transportes y Comunicaciones		
10	2018	55	634	Condiciones de Vida y Pobreza - ENAH0	11	Servicios a la Vivienda		

Especialidades mas estudiadas por los limeños



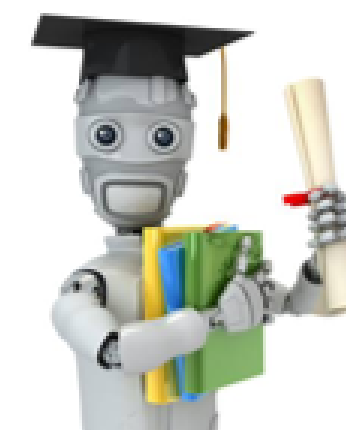
Fuente: INEI

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"

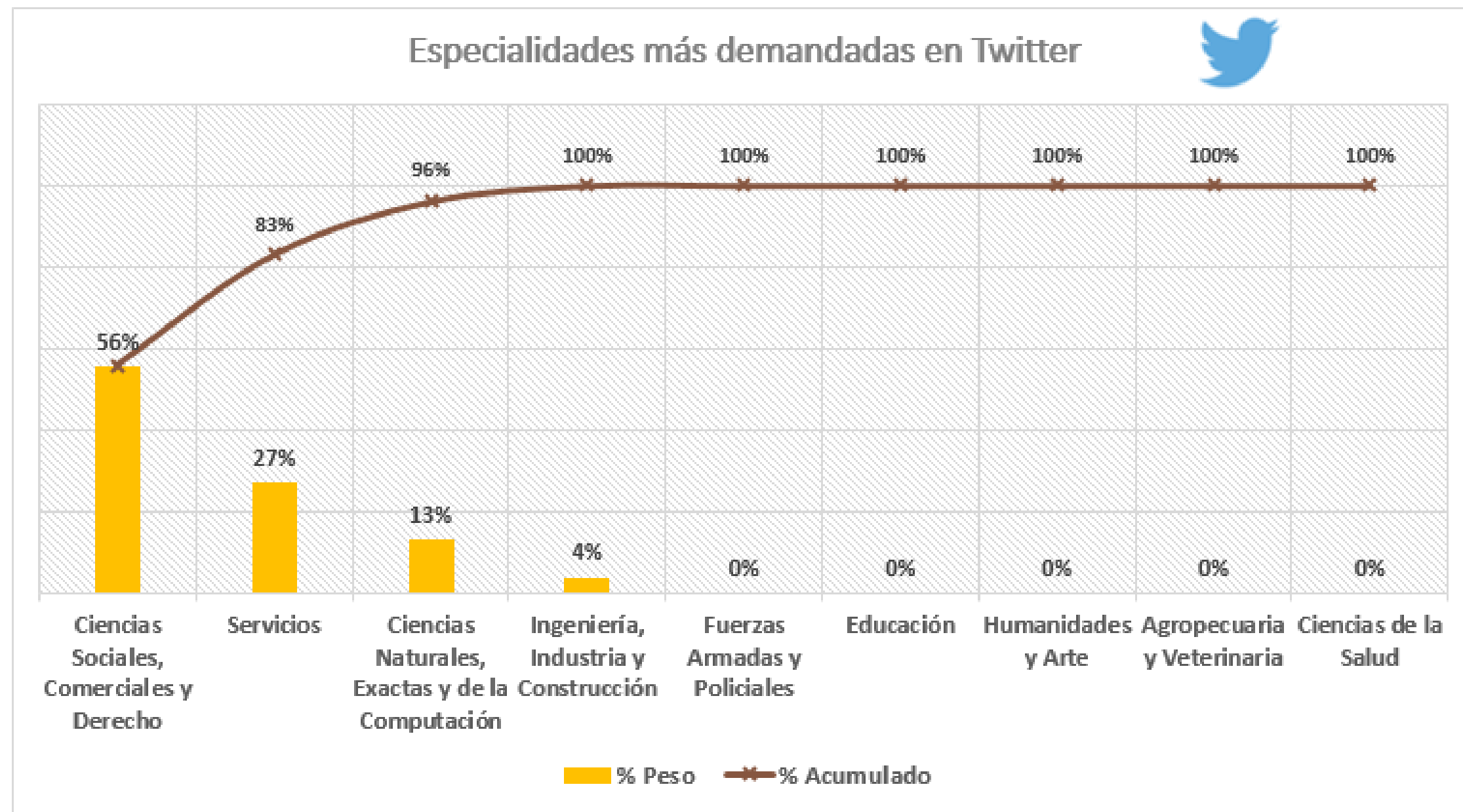
Topic: 6 Word: 0.023*"perú" + 0.017*"empleo" + 0.016*"para" + 0.015*"empresa" + 0.014*"transna" + 0.010*"overal" + 0.009*"trabajo" + 0.008*"reclutamiento" + 0.008*"industri" + 0.008*"tien"

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"

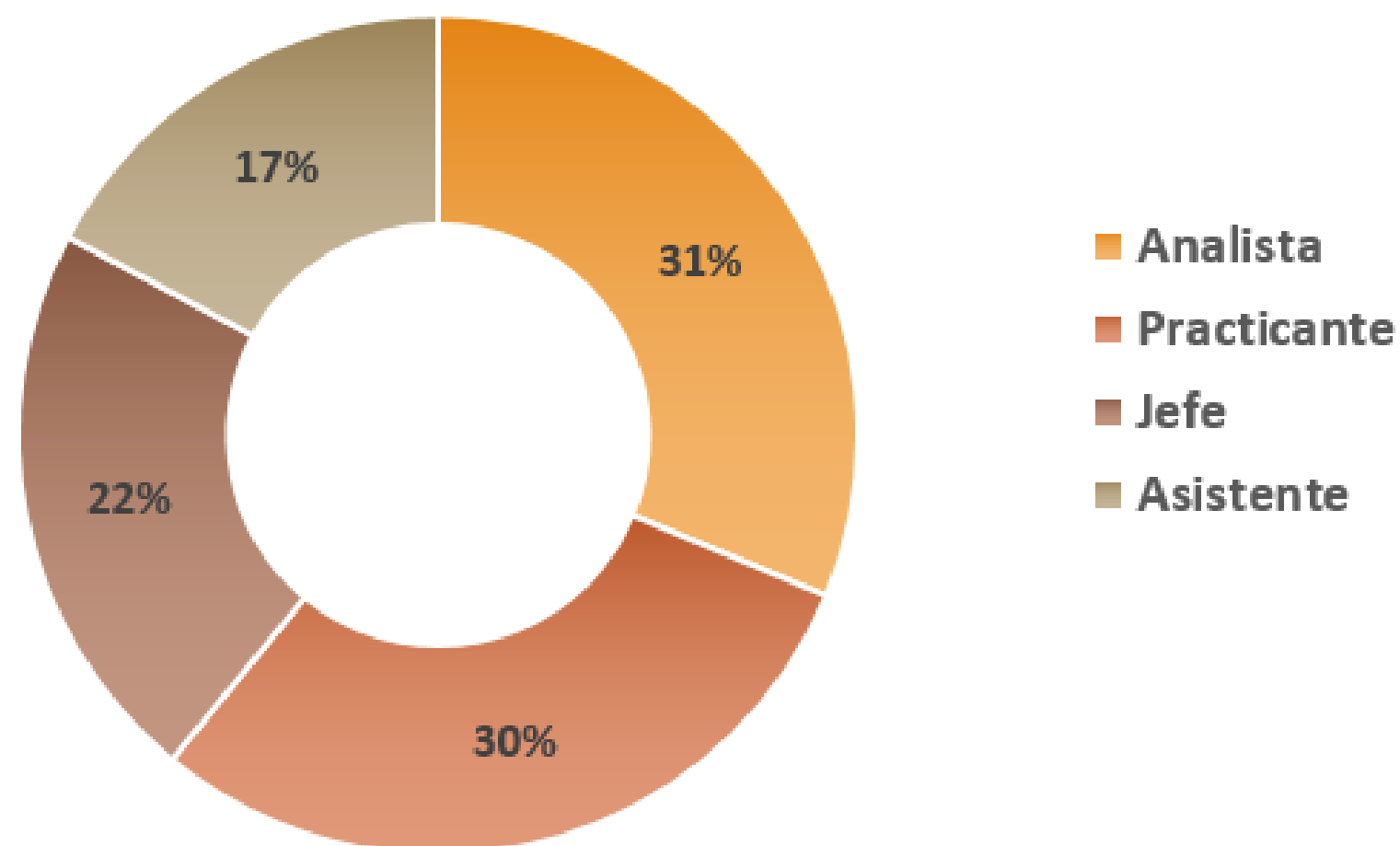
Topic: 9 Word: 0.016*"perú" + 0.016*"empresa" + 0.016*"client" + 0.016*"programador" + 0.015*"nuestro" + 0.014*"jefe" + 0.012*"empleo" + 0.012*"venta" + 0.011*"servicio" + 0.010*"analista"

Especialidades mas demandadas en Twitter



Cargos mas demandados en Twitter

Cargos más demandados en Twitter



Ingesta y Almacenamiento

Compute Engine

Instancias de VM

CREAR INSTANCIA IMPORTAR VM ACTUALIZAR INICIAR DETENER RESTABLECER

Filtrar las instancias de VM

Columnas

<input type="checkbox"/>	Nombre	Zona	Recomendación	Usada por	IP interna	IP externa	Conectar
<input type="checkbox"/>	<input checked="" type="checkbox"/> cluster-comiters-m	southamerica-east1-a			10.158.0.3 (nic0)	35.198.77.71	SSH
<input type="checkbox"/>	<input checked="" type="checkbox"/> cluster-comiters-w-0	southamerica-east1-a			10.158.0.4 (nic0)	35.198.36.138	SSH
<input type="checkbox"/>	<input checked="" type="checkbox"/> cluster-comiters-w-1	southamerica-east1-a			10.158.0.2 (nic0)	35.247.221.161	SSH
<input type="checkbox"/>	<input checked="" type="checkbox"/> cluster-e9d5-m	europa-west1-b			10.132.0.2 (nic0)	Ninguna	SSH
<input type="checkbox"/>	<input checked="" type="checkbox"/> cluster-e9d5-w-0	europa-west1-b			10.132.0.4 (nic0)	Ninguna	SSH
<input type="checkbox"/>	<input checked="" type="checkbox"/> cluster-e9d5-w-1	europa-west1-b			10.132.0.3 (nic0)	Ninguna	SSH

```
comiters-m:~$ hdfs dfs -ls -R comiters
Mijail hadoop      0 2019-06-30 19:35 comiters/datalake
Mijail hadoop      0 2019-07-06 09:20 comiters/datalake/DDA
Mijail hadoop      0 2019-07-06 09:07 comiters/datalake/DDA/Twitter
Mijail hadoop    452 2019-07-06 09:07 comiters/datalake/DDA/Twitter/Especialidades_mas_demandadas_en_Twitter.csv
Mijail hadoop      0 2019-07-06 09:21 comiters/datalake/DDA/Twitter2
Mijail hadoop    156 2019-07-06 09:21 comiters/datalake/DDA/Twitter2/Cargos_mas_demandados_en_Twitter.csv
Mijail hadoop      0 2019-06-30 19:36 comiters/datalake/RDA
Mijail hadoop      0 2019-07-06 11:31 comiters/datalake/RDA/educacion
Mijail hadoop  1133817 2019-07-06 11:31 comiters/datalake/RDA/educacion/1_Tabla_Educación+Llave.csv
Mijail hadoop   845424 2019-07-06 11:31 comiters/datalake/RDA/educacion/2_Tabla_Empleo+Llave.csv
Mijail hadoop      0 2019-07-06 10:54 comiters/datalake/RDA/empleo
Mijail hadoop      0 2019-06-30 19:35 comiters/datalake/UDA
comiters-m:~$
```


Ingesta y Almacenamiento

```
#-----#
# CREACION DE TABLAS #
#-----#

# Educación
#-----#
use comiters;
CREATE EXTERNAL TABLE IF NOT EXISTS comiters.TablaEducacion(
ano string COMMENT 'AÑO',
Llave string COMMENT 'LLAVE',
conglome string COMMENT 'CONGLOME',
vivienda string COMMENT 'VIVIENDA',
hogar string COMMENT 'HOGAR',
codperso string COMMENT 'CODPERSO',
ubigeo string COMMENT 'UBIGEO',
Distrito string COMMENT 'DISTRITO',
grado_de_estudio string COMMENT 'GRADO_ESTUDIO',
carrera string COMMENT 'CARRERA',
especialidad string COMMENT 'ESPECIALIDAD',
edad int COMMENT 'EDAD'
)
COMMENT 'Tabla educacion'
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ';'
LINES TERMINATED BY '\n'
STORED AS TEXTFILE
LOCATION '/comiters/datalake/DDA'
tblproperties("skip.header.line.count" = "1");
```

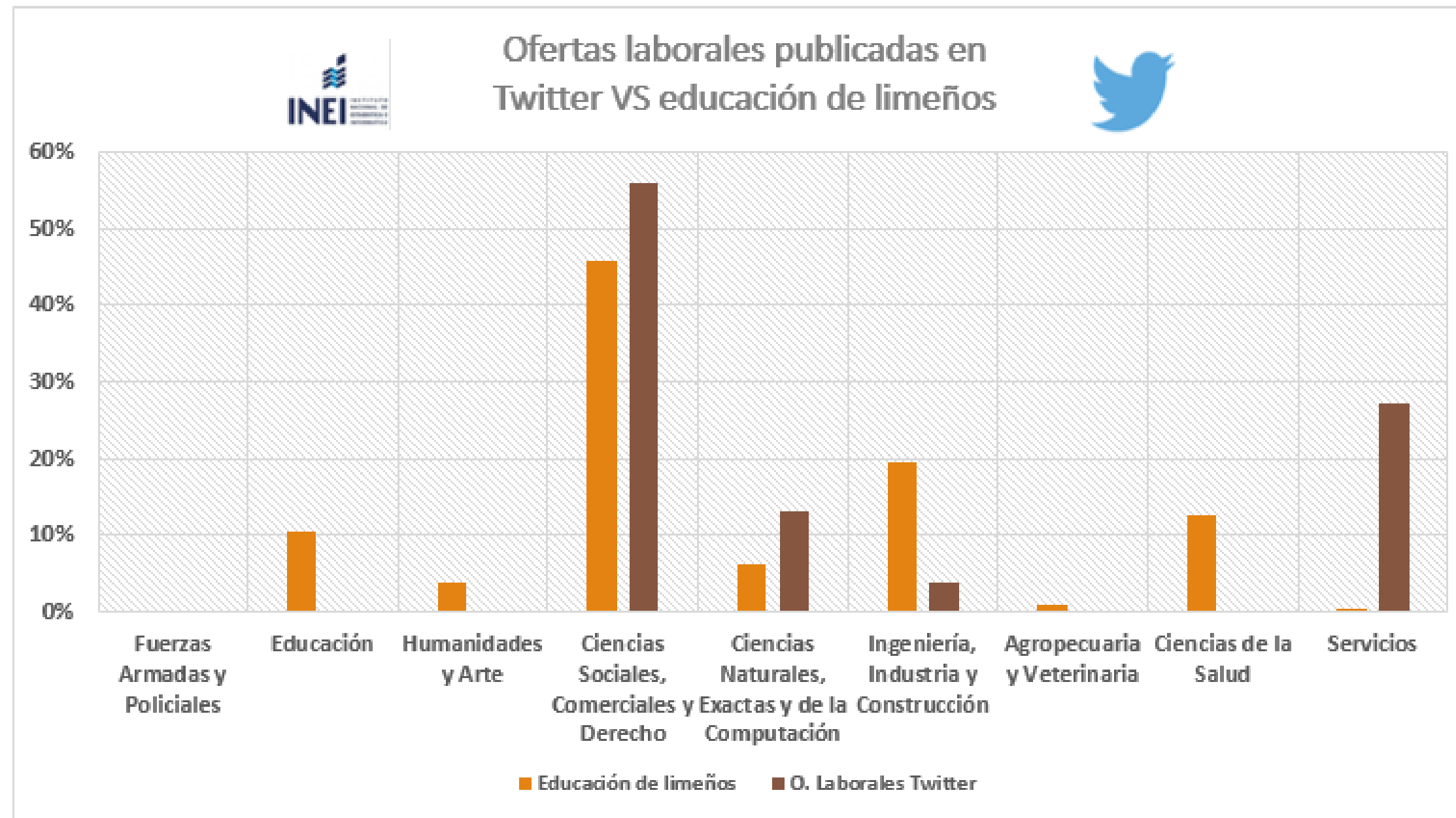
```
# CARGANDO LAS TABLAS
load data inpath 'comiters/datalake/RDA/educacion/1_Tabla_Educación+Llave.csv' into table comiters.TablaEducacion;
# CARGANDO LAS TABLAS
load data inpath 'comiters/datalake/RDA/empleo/2_Tabla_Empleo+Llave.csv' into table comiters.TablaEmpleo;
```

```
# Empleo
#-----#
use comiters;
CREATE EXTERNAL TABLE IF NOT EXISTS comiters.TablaEmpleo(
ano string COMMENT 'AÑO',
Llave string COMMENT 'LLAVE',
conglome string COMMENT 'CONGLOME',
vivienda string COMMENT 'VIVIENDA',
hogar string COMMENT 'HOGAR',
codperso string COMMENT 'CODPERSO',
ubigeo string COMMENT 'UBIGEO',
Distrito string COMMENT 'DISTRITO',
genero string COMMENT 'GENERO',
situacion_laboral string COMMENT 'SITUACIÓN LABORAL',
condicion_de_empleo string COMMENT 'CONDICIÓN DE EMPLEO',
pago_mensual int COMMENT 'EDAD'
)
COMMENT 'Tabla empleo'
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ';'
LINES TERMINATED BY '\n'
STORED AS TEXTFILE
LOCATION '/comiters/datalake/DDA'
tblproperties("skip.header.line.count" = "1");
```

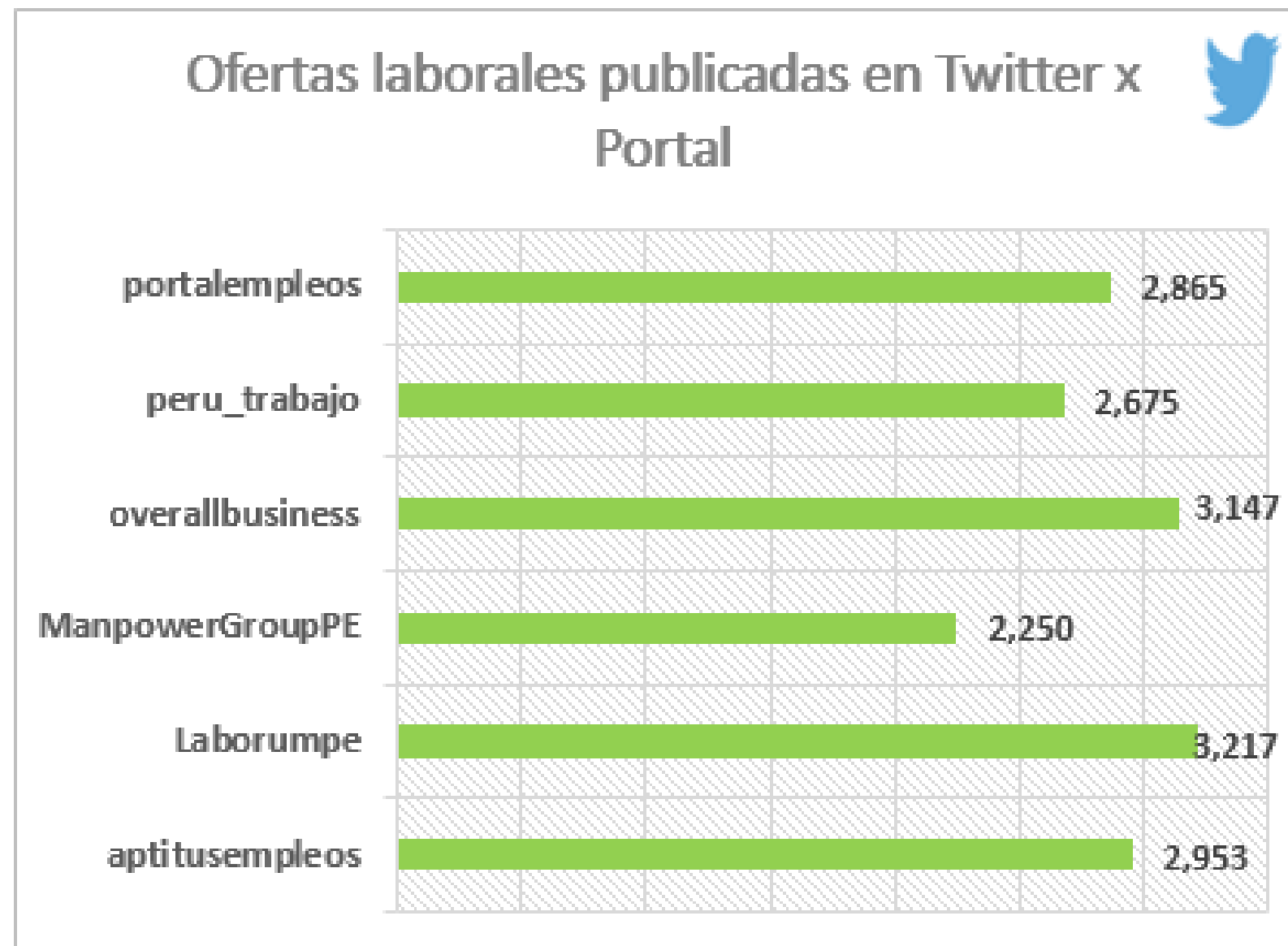
Ingesta y Almacenamiento

```
Time taken: 0.108 seconds, Fetched: 2 row(s)
hive> use comiters;
OK
Time taken: 0.04 seconds
hive> describe comiters.TablaEducacion;
OK
ano                string      AÑO
llave              string      LLAVE
conglome           string      CONGLOME
vivienda           string      VIVIENDA
hogar              string      HOGAR
codperso           string      CODPERSO
ubigeo             string      UBIGEO
distrito           string      DISTRITO
grado_de_estudio   string      GRADO_ESTUDIO
carrera            string      CARRERA
especialidad       string      ESPECIALIDAD
edad              int         EDAD
Time taken: 0.095 seconds, Fetched: 12 row(s)
hive> describe comiters.TablaEmpleo;
OK
ano                string      AÑO
llave              string      LLAVE
conglome           string      CONGLOME
vivienda           string      VIVIENDA
hogar              string      HOGAR
codperso           string      CODPERSO
ubigeo             string      UBIGEO
distrito           string      DISTRITO
genero             string      GENERO
situacion_laboral  string      SITUACIÓN LABORAL
condicion_de_empleo string      CONDICIÓN_DE_EMPLEO
pago_mensual       int         EDAD
Time taken: 0.055 seconds, Fetched: 12 row(s)
hive> select * from comiters.TablaEducacion;
OK
2018  602616111  6026  16  11  1  70106  Ventanilla  secundar  37
2018  602616112  6026  16  11  2  70106  Ventanilla  primaria  29
2018  602616113  6026  16  11  3  70106  Ventanilla  primaria  10
2018  602616114  6026  16  11  4  70106  Ventanilla  primaria  7
2018  602616115  6026  16  11  5  70106  Ventanilla  sin nive  3
2018  602643111  6026  43  11  1  70106  Ventanilla  superior  electricidad  Ingenieria, Industria y Construcción 49
2018  602643112  6026  43  11  2  70106  Ventanilla  superior  disenio de m  Ingenieria, Industria y Construcción 21
2018  602643113  6026  43  11  3  70106  Ventanilla  secundar  13
```

4. Visualización de Resultados



Sizing de la generación de data

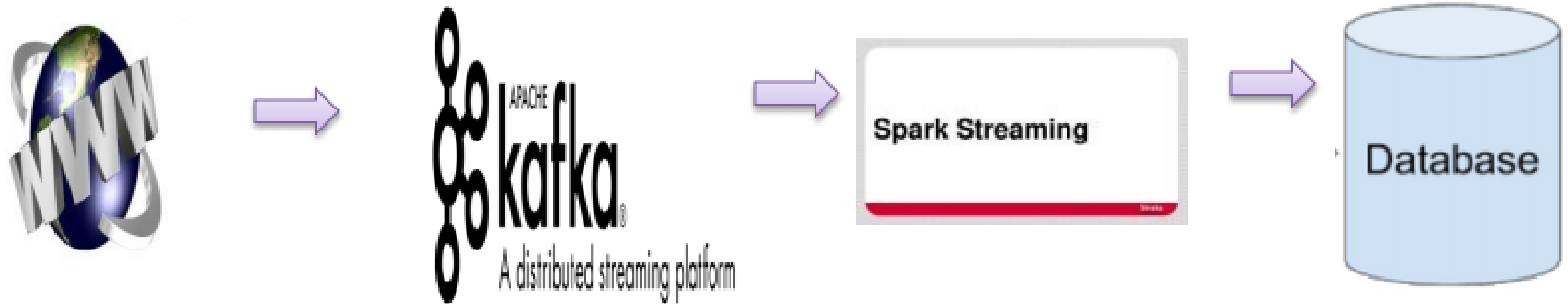


La data obtenida entre los años 2012 y 2018 fue:

Promedio	Cantidad
Mensual	201
Día	7
Hora	0,27
Minuto	0,0046

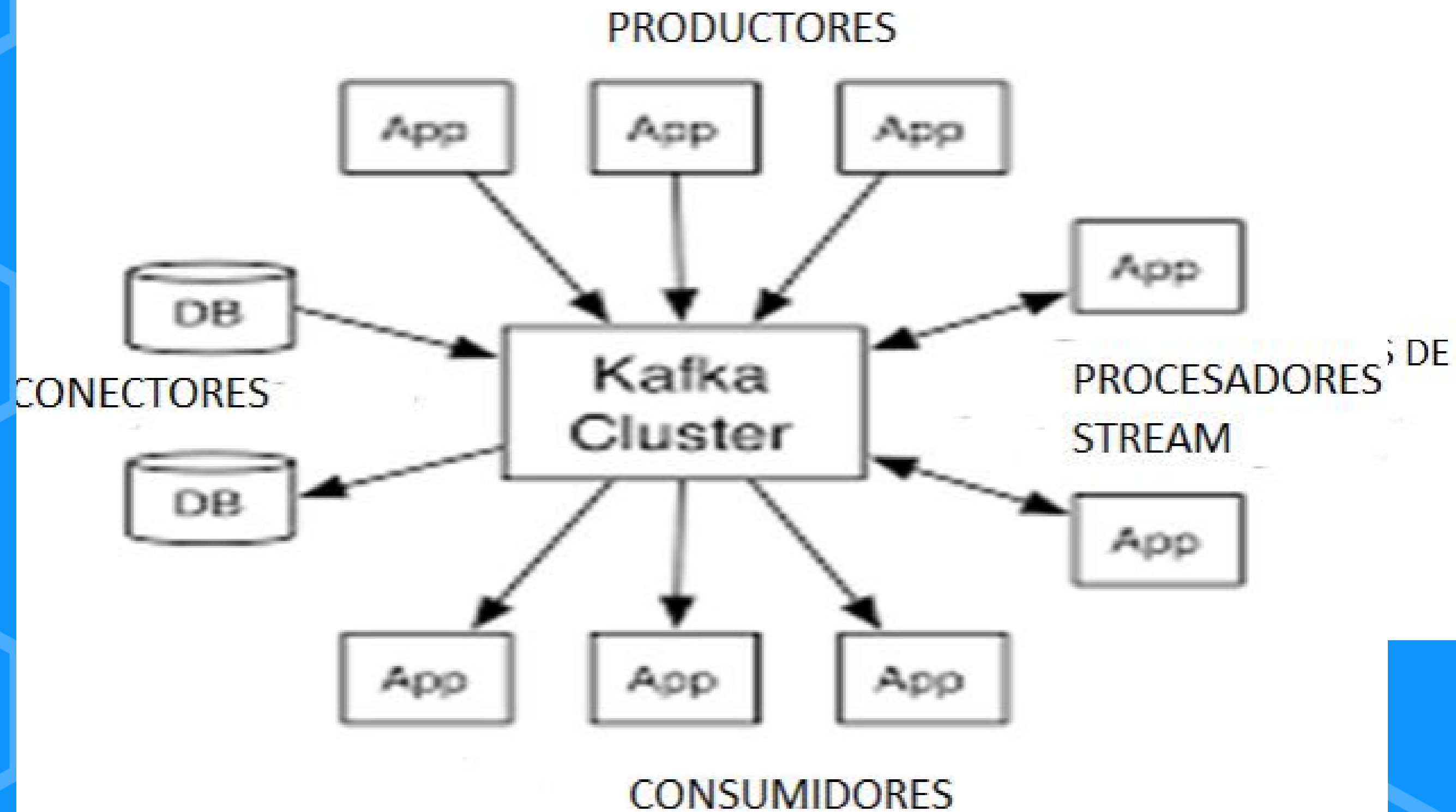
ARQUITECTURA STREAMING CON KAFKA

A nivel de diagrama de Caja Negra



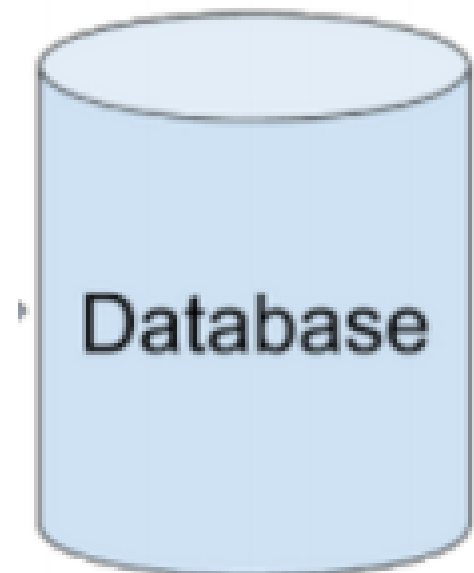
ARQUITECTURA STREAMING CON KAFKA

A nivel de caja blanca



Arquitectura Streaming con Apache Flume

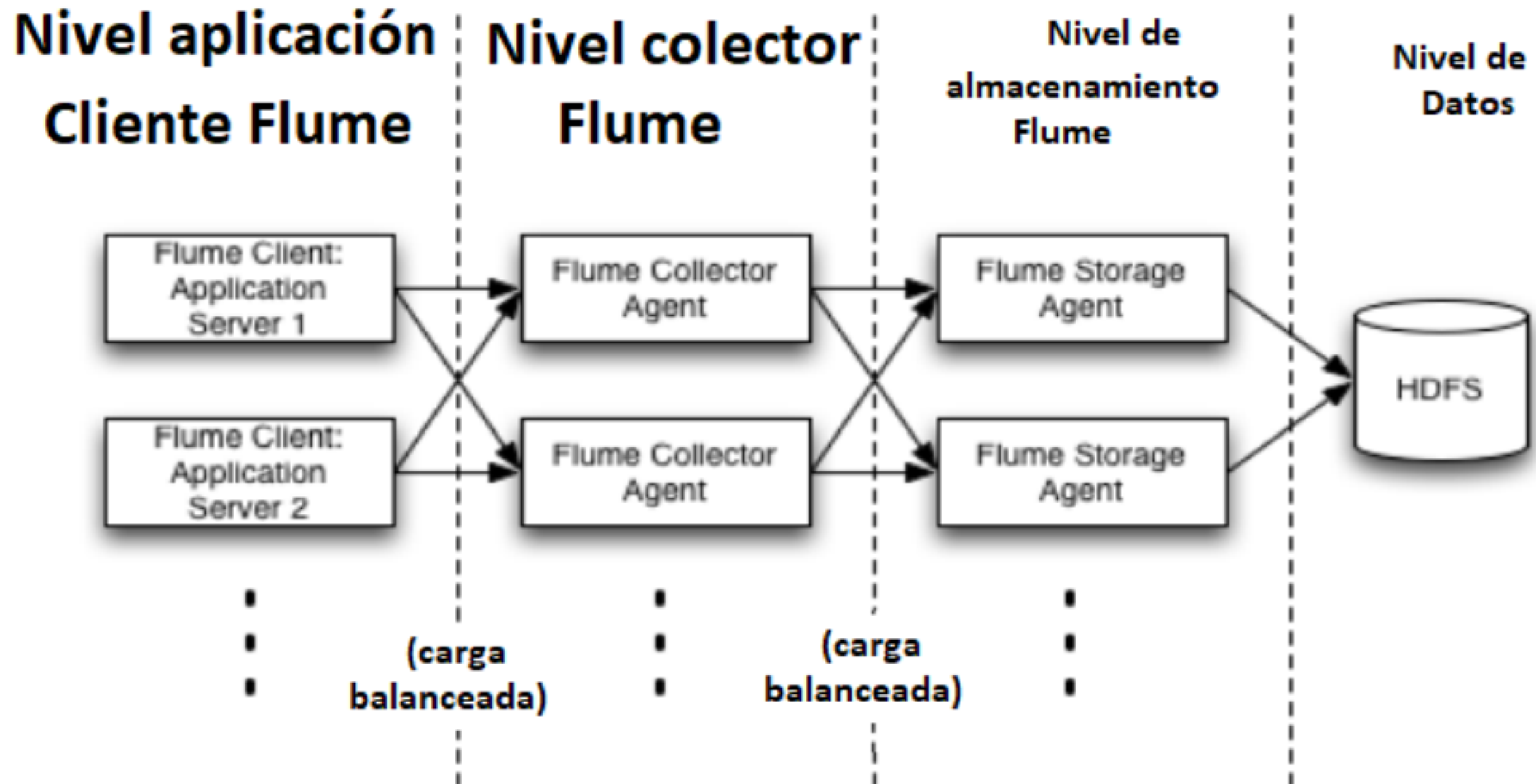
A nivel de caja negra



Arquitectura Streaming con Apache Flume

A nivel de diagram de caja blanca

Una topología de bloques Flume



Anexos:

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

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 artículo en Medium	Eduardo			x

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 paginas 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 metricas e indicadores a construirse, ademas de justificar la aplicación del Big Data (3Vs)	Eduardo			x

GRACIAS TOTALES



The background of the slide features a repeating pattern of light blue hexagons on a darker blue background. This pattern is visible in the top and bottom sections of the slide, framing a central dark purple band.

MUCHAS GRACIAS

CTIC