



Grupo 4

Proyecto Final Análisis
demográfico mediante la
interacción en redes sociales,
para la optimización de
hotspots municipales en la
ciudad de Quito

Andrade Rodrigo
Pinchao Alexander
Polanco Boris
Sosa Luis

1 INTRODUCCION

El uso de herramientas que permiten el análisis de grandes cantidades de datos, con la finalidad de extraer información útil para la generación de estrategias; toman una vital importancia dentro de las sociedades actuales. Estas fuentes de datos pueden ser variadas, pero en casos en los cuales se desea determinar el comportamiento de los individuos; es preciso usar métodos que nos permitan la recolección de datos de manera no invasiva. Con esto nos referimos a realizar un estudio de su comportamiento sin recurrir a la observación directa. Con esto nos aseguramos el respeto al derecho a la privacidad de los individuos y evitamos sesgos en el estudio debidos a prejuicios.

Es preciso seleccionar la fuente de datos para poder realizar un análisis, fruto del cual se podrá entender el comportamiento de los individuos y la forma en la cual se podrá mejorar su entorno. En el caso de nuestro estudio es preciso recolectar información, de manera que podamos determinar sus costumbres y hábitos; con esto en mente analizamos la interacción en redes sociales. Con esto obtendremos los datos necesarios para poder desarrollar una solución, para optimizar la distribución de puntos de acceso municipales.

2 PERIODO DE DESARROLLO

Las habilidades para el desarrollo del proyecto han sido desarrolladas en el transcurso de 2 meses, dentro del marco del curso Big Data y Data Analytics, mientras que los datos expuestos han sido recogidos a lo largo de 5 semanas

3 OBJETIVOS

- Procesar los datos obtenidos con la finalidad de generar información, útil para el desarrollo de estrategias de utilidad para el DMQ (Distrito Metropolitanos de Quito).
- Aplicar las destrezas aprendidas en el desarrollo del curso Big Data y Data Analytics para desarrollar un conjunto de indicadores de la situación actual.

4 ALCANCE

Se realizara el análisis de las interacciones de individuos dentro de Twitter (servicio de microblogging), en el DMQ; estas se tomaran en un boundigbox que se enmarca en la zona urbana de la ciudad.

5 DEFINICION DEL PROBLEMA

La correcta gestión de los recursos públicos implica la generación de proyectos, que tengan un impacto positivo en la sociedad; pero esto no es suficiente para considerar la gestión como exitosa.

Uno de los pasos más importantes en el ciclo de vida de un proyecto, es el control de los resultados y la eventual mejora del mismo. Un proyecto de gran impacto como es la creación de hotspots municipales, el cual ha demostrado su éxito con la implementación de 21 zonas (84 puntos de acceso) WiFi; con un funcionamiento continuo.

El mantener estos puntos de acceso resulta costoso, más aun si tenemos en cuenta que estos puntos no tendrán siempre una demanda constante; es decir en determinados momentos estos se encontraran sub-utilizados. Es en este contexto donde el uso de una herramienta que permita analizar el comportamiento de los individuos, resulta de vital importancia; ya que nos permitirá optimizar los recursos y analizar la posible expansión de la red.

6 JUSTIFICACION

La interacción de individuos en redes sociales genera cantidades de datos, los cuales tienen varias dificultades para su procesamiento, estas dificultades van a ser expuestas valiéndose del principio de las 5 v de big data.

6.1 Variedad

El flujo de datos será en su mayor parte de tipo semiestructurado y no estructurados los cuales serán obtenidos mediante la API de las redes sociales y diferentes métodos de minería y recolección.

6.2 Volumen

La captura continuada de tweets genera ingentes cantidades de información, la cual crece constantemente a un ritmo de 50 TPS, guardando la información en modo texto plano (200bytes) son 10K por segundo, 600K por minuto, 36 Megas por hora, 864 megas, al día.

6.3 Velocidad

Los datos generados por la interacción en redes sociales arriban a gran velocidad, lo cual genera que estos se acumulen en un corto periodo de tiempo, y no puedan ser procesados correctamente y así obtener un valor.

6.4 Veracidad

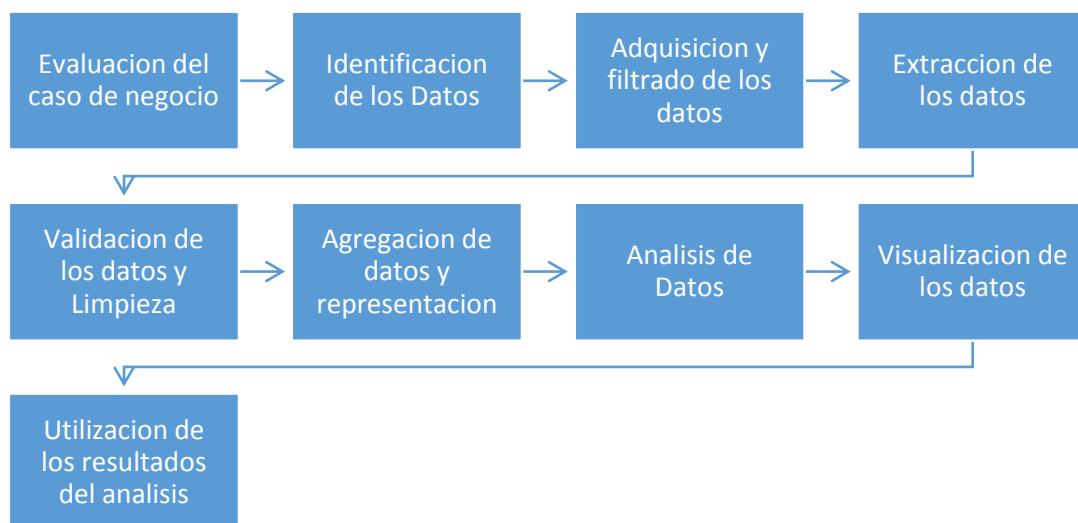
Si definimos la veracidad como la calidad o fidelidad de los datos, los datos obtenidos tendrán un contenido alto de ruido, es decir datos que no serán útiles para el análisis a ser realizado, por lo tanto estos deberán ser depurados

6.5 Valor

Se busca optimizar los hotspots municipales ubicándolos en áreas donde sea mayor el número de usuarios potenciales y gestionar el periodo de actividad de los mismos para reducir la utilización de recursos.

7 DESARROLLO

Para el desarrollo de una solución que nos permita obtener los resultados deseados, se siguió la estructura del ciclo de vida de la analítica de Big Data.



7.1 EVALUACIÓN DEL CASO DE NEGOCIO

La generación de una solución que permita el análisis de los datos para el desarrollo de este proyecto, hace que la conformación de un equipo multidisciplinario sea necesario.

Entre los profesionales que conforman el equipo de trabajo, podemos encontrar profesionales con variados perfiles, entre los que tenemos:

Integrantes	Áreas de destreza
Andrade Rodrigo	Experto en bases de datos
Pinchao Alexander	Desarrollador de software
Polanco Boris	Ingeniero matemático
Sosa Luis	Experto Business intelligence

Como recursos para la realización del proyecto usaremos:

Recursos	Coste
Bases de Datos	
Couchdb	0
Smileupps (DBaaS couchdb)	0 (Dentro de periodo de prueba)
Indexación y análisis de datos	
Elasticsearch 1.7.5	0
Plugins Elasticsearch	
Kibana 4.0.1	0
River	0
Head	0
Kopf	0
Lenguajes de programación	
R	0
Python 2.7	0
javascript	0
Equipos	
Computadores Dell i7 del laboratorio sigma	0
MacBook Pro i5	0
Laptop Dell Pentium 4	0
Capacitación	
Sin capacitación aparte del curso	0
Total	0

7.2 IDENTIFICACIÓN DE LOS DATOS

El estudio de la interacción de individuos en una sociedad, nos propone una serie de obstáculos. Para el desarrollo de este proyecto podemos encontrar entre los más importantes, los de índole legal definida como los derechos inalienables de las personas, y los del ámbito propio al estudio.

Entre los de índole legal podemos citar el derecho a la privacidad y entre los de ámbito propio al estudio; el propio comportamiento (o cambio del mismo) de los sujetos de estudio; al ser conocido el hecho de estar bajo estudio.

Por lo tanto es necesario escoger una plataforma que nos permita recoger información, sin que sea conocido el hecho de que lo estamos realizando y que permita al sujeto de estudio permanecer en anonimato; esto para asegurarnos las condiciones de doble-ciego necesarias para la conducción de un estudio riguroso.

Por lo tanto es necesario una herramienta con una API para la comunicación, que maneje su información con seudónimos y con un grado de penetración elevado.

Herramienta	Privacidad	API	penetración	Velocidad de recolección	subjetividad
Encuesta personal			X		
Facebook		X	X	X	X
Instagram		X		X	X
Linkedin		X		X	
Twitter	X	X	X	X	X

7.3 ADQUISICIÓN Y FILTRADO DE DATOS

El uso de esta expresión regular nos permite el eliminar datos incoherentes, como podrían ser “ubicación”: “Tres metros sobre el cielo” o “Ciudad”: “Carita de dios”. Además de permitirnos el definir exactamente el lugar de recolección.

Además se realizó el proceso de respaldo de la información a una base de datos en la nube, con el comando

- `curl -H 'Content-Type: application/json' -X POST http://localhost:5984/_replicate -d '{"source": "http://127.0.0.1:5984/quito", "target": "https://couchdb-87ce75.smileupps.com/quito", "create_target": true, "continuous": true}' [2]`

7.4 EXTRACCION DE LOS DATOS

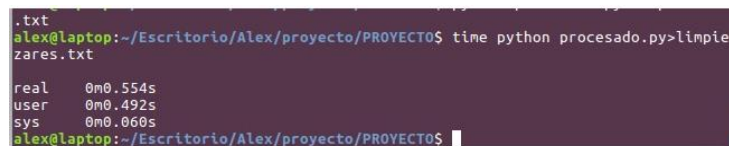
Para la siguiente fase del desarrollo de la solución se prosiguió a descargar el contenido de la base de datos a un archivo json usando el comando, (el procesamiento de la información así como el desarrollo de las soluciones fueron realizadas sobre una plataforma Linux Ubuntu 16.04):

- `curl -X GET http://127.0.0.1:5984/BaseDeDatos/_all_docs?include_docs=true > BaseDeDatos.json [2]`

Una vez obtenido el archivo son este es analizado para determinar cuáles de ellos poseen información necesaria con el uso de una expresión regular.

- `created_at":([^\,]+),"f[^"]*"([^\,]+)([^\,]+) [3]`

En la cual el primer grupo de paréntesis recoge los valores de las coordenadas del boundingbox y el segundo la fecha de la creación; usando estas expresiones el tiempo de procesamiento se reduce significativamente logrando procesar 40000 tweets en 0.5 segundos.



```
.txt
alex@laptop:~/Escritorio/Alex/proyecto/PROYECTO$ time python procesado.py > limpiezares.txt
real    0m0.554s
user    0m0.492s
sys     0m0.060s
alex@laptop:~/Escritorio/Alex/proyecto/PROYECTO$
```

Los archivos son impresos en consola pero pueden ser re-direccionados a un archivo de texto como se ve en la imagen.

7.5 VALIDACIÓN Y LIMPIEZA DE DATOS

Al obtener datos directamente desde un archivo json, estos contendrán símbolos que se encuentran dentro del lenguaje de marcado es decir comillas, paréntesis, llaves, corchetes y comillas dobles. Estos caracteres especiales pueden alterar la naturaleza de nuestros datos por lo cual es necesario el eliminarlos de los datos, para este fin usamos las funciones propias de Python. [4] [3]

Una vez hecho esto consolidamos los datos de coordenadas (4 puntos) en un solo punto [5].

7.6 AGREGACIÓN Y REPRESENTACIÓN DE DATOS

Los datos una vez validados y limpiados son enviados a la base de datos, transformándolos a objetos json. [6] Una vez hecho esto se los datos son replicados a una base de datos en la nube, desde todos los colectores, a la base de datos (https://couchdb-87ce75.smileupps.com/_utils/database.html?coordenadas).

- `curl -H 'Content-Type: application/json' -X POST http://localhost:5984/_replicate -d '{"source": "http://127.0.0.1:5984/quito", "target": "https://couchdb-87ce75.smileupps.com/coordenadas", "create_target": true, "continuous": true}'`

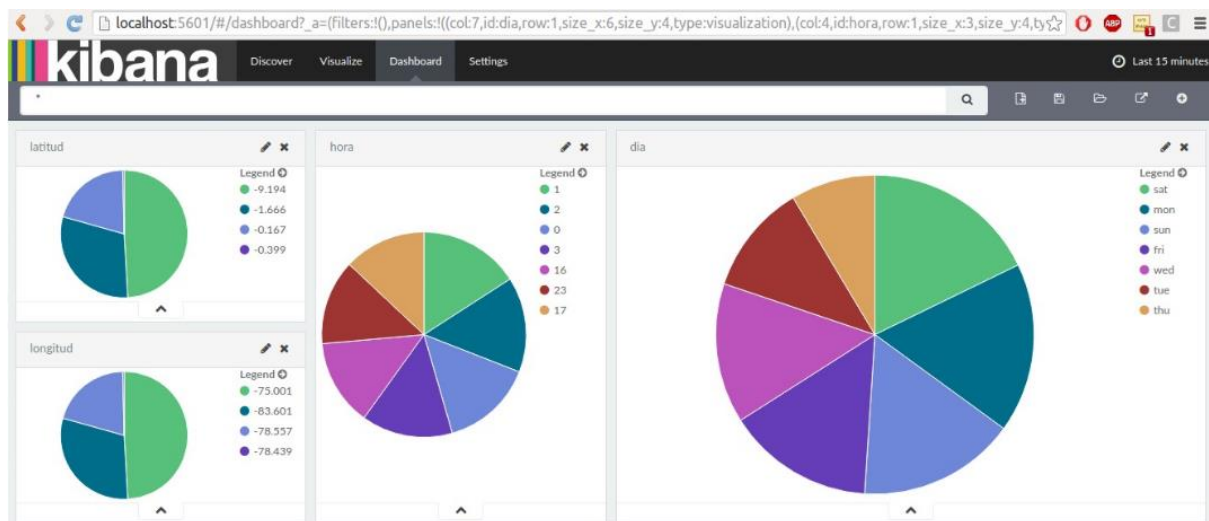
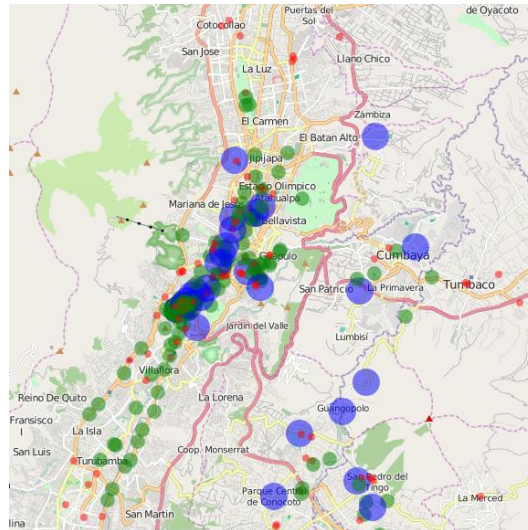
7.7 Análisis de los datos

Para el análisis de la información, se realizó de primera mano un análisis de datos de confirmación, ya que intuíamos que las zonas con mayor actividad serían aquellas dentro de la zona urbana, fuera de los horarios de oficina, pero al final esto resultó ser verdad solo en parte.

Ya que se encontraron comportamientos interesantes que determinan un patrón de conducta recurrente de los individuos.

7.8 Visualización de los datos

Para la visualización de los datos utilizamos kibana para el análisis del comportamiento dentro de los días y las horas del día de mayor actividad y para las coordenadas utilizamos R [7].

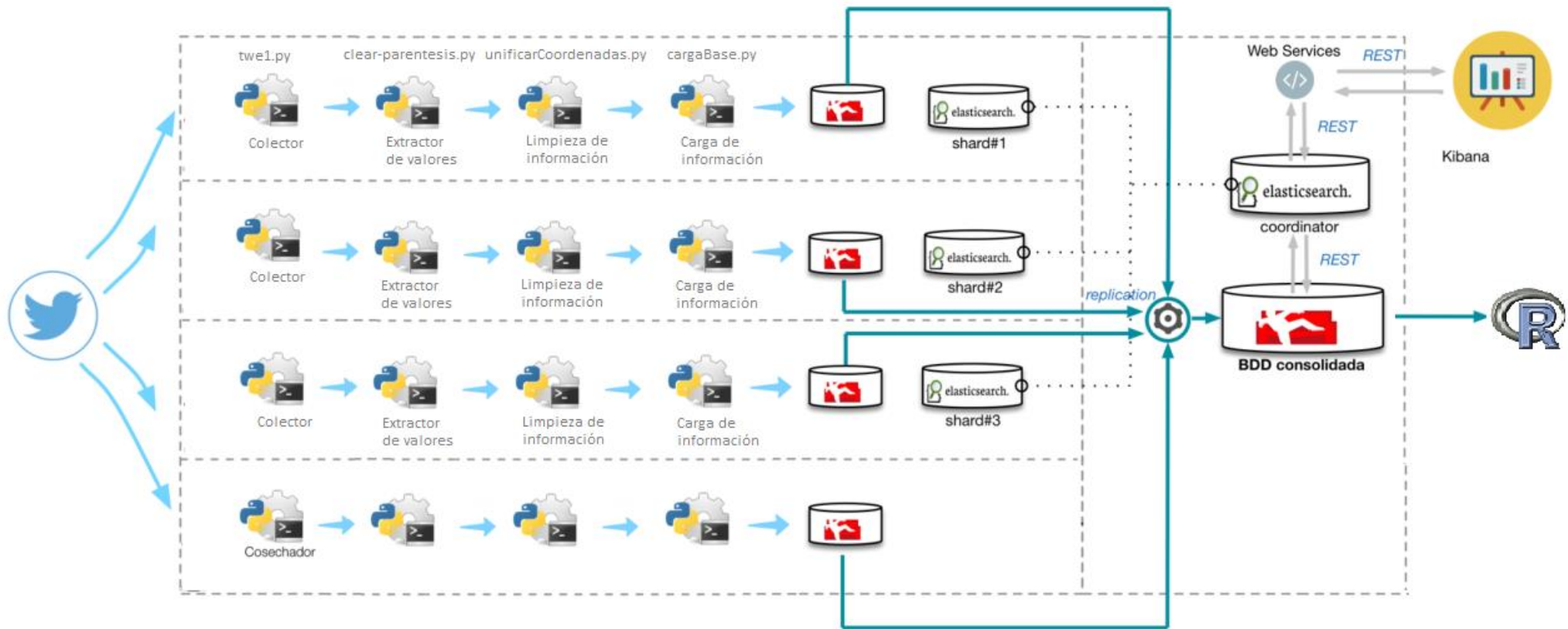


7.9 Utilización de los resultados

Los datos obtenidos podrían ser utilizados por la alcaldía del DMQ para mejorar y ampliar las zonas de acceso inalámbrico. Ya que existen lugares en los cuales el tráfico es intensivo, pero no cuentan con este beneficio (como por ejemplo Guapulo), además podría reducir los valores del ancho de banda o directamente apagar los hotspots que no estén utilizando para transmitir este ancho de banda a los que si los necesitan.

8 Arquitectura de la solución

8.1 Grafica de la arquitectura



8.2 Explicación de la arquitectura

8.2.1 Recolección de la información

Para la recolección de la información se utilizó un script que permitía acceder a la API de twitter a cada uno de los miembros del grupo, en función a los valores del boundingbox determinado para cada miembro y las credenciales de acceso a twitter [8]. Cada miembro del grupo realizaba la recolección de datos en una base de datos local en sus propios equipos.

8.2.2 Procesamiento de la información

Los scripts para la extracción de texto, limpieza de caracteres especiales, cálculo de coordenadas y carga a la base de datos se ejecutaban en cada una de las maquinas.

8.2.3 Unificación de la información

Una vez procesada la información esta es enviada a una base de datos en la cual la información era consolidada en una máquina que haría las veces de maquina principal.

8.2.4 Indexación

Una vez consolidada la información en una sola maquina se armó el clúster de 3 máquinas las cuales tenían 3 shards, para su indexación siendo la maquina en la que se consolido la información la misma máquina que haría las veces de master del clúster para el procesamiento de la información.

8.2.5 Visualización

Para la visualización se instaló Kibana en la misma maquina master para manejar los datos de tipo String y los datos doublé con R desde la misma base de datos.

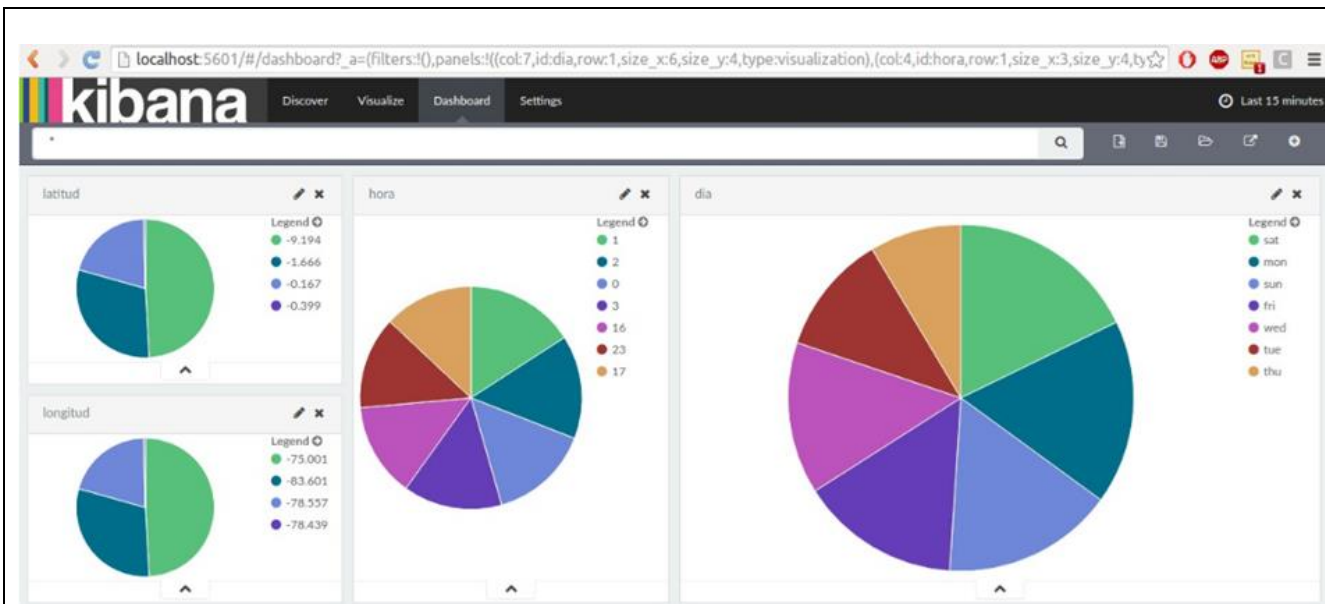
9 RESULTADOS OBTENIDOS

Los resultados obtenidos del análisis indican que:

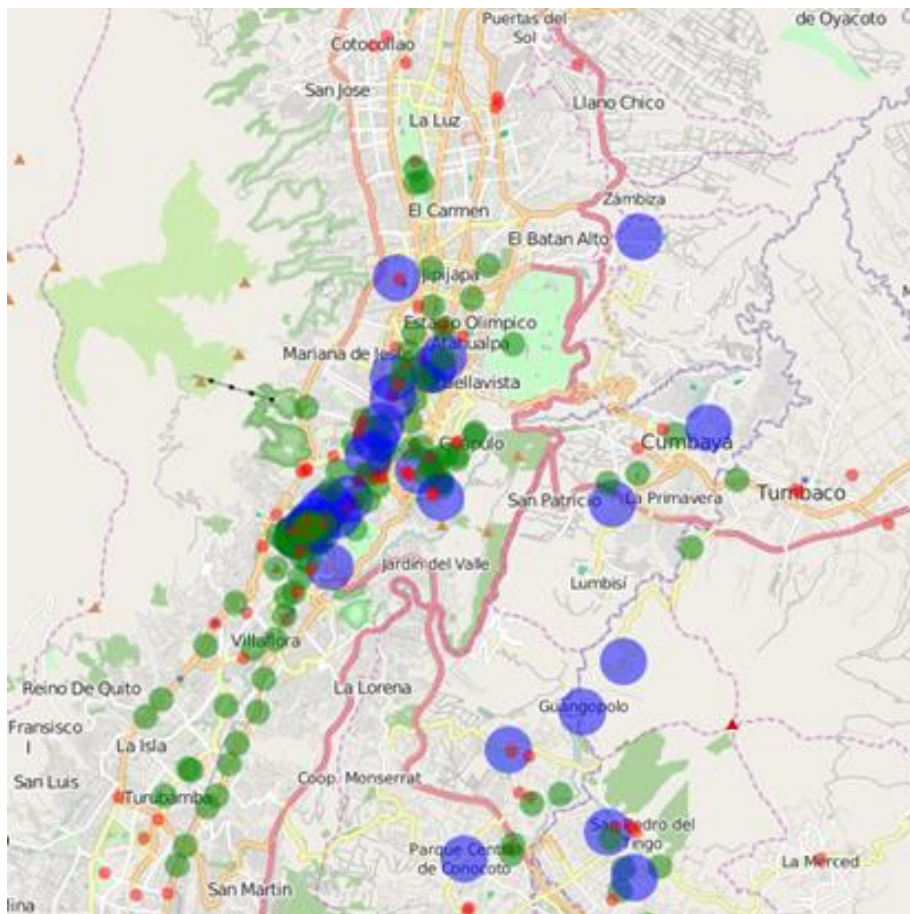
- Los días en los cuales las personas envían tweets con mayor frecuencia son los días sábados,
- Los horarios con mayor frecuencia de tweets son entre las 12 y las 3 de la madrugada
- Los sitios en los que más tweets se generan son cercanos a las áreas en las cuales se encuentran mayor número de discotecas y bares.

De estos 3 resultados podemos inferir que los horarios de mayor tráfico en twitter son los días sábados en la zona rosa de quito (Área que comprende la plaza Foch, La mariscal, y Guapulo), lo que nos indica que el estudio esta sesgado, ya que esto nos indica que la mayor parte de los usuarios son, personas mayores de edad, los cuales podrían ser considerados adulto joven [9].

Del análisis además se concluyó que solo una fracción de los tweets pueden ser usados para este tipo de estudios, esto debido a que la mayor parte de usuarios no le conceden acceso a la aplicación de twitter para acceder a los datos de GPS; lo que provoca que las coordenadas del tweet sean las correspondientes a las antenas de re-transmisión en cruz loma.



Las horas y los días de mayor actividad en twitter



Distribucion de los tweets generados en la área de quito

10 Problemas encontrados y soluciones

Problema	Solución
Tiempo de carga a la base de datos demasiado largo luego del proceso de limpieza y filtrado de la información	Realizar la limpieza y filtrado de la información en tiempo de ejecución, para poder realizar un solo ingreso.
Pocas ocurrencias de usuarios con GPS activado	Tomar muestras por un periodo más largo, con la finalidad de aumentar el número de tweets útiles para el análisis.
Fallos en la conexión de Elasticsearch nodos y master	Revisar los firewalls del sistema operativo y los de los antivirus
Un nodo no se reconoce en el cluster generado	Revisar la configuración y la versión de los nodos

11 Conclusiones y recomendaciones

11.1 Conclusiones

- Las soluciones de big Data a ser desarrolladas deben ser planificadas de manera exhaustiva, ya que el menor descuido podría implicar la reducción de la eficiencia de la misma.
- Existen gran variedad de herramientas, que nos sirven en las distintas instancias del ciclo de vida de los proyectos de big Data, pero es responsabilidad del analista el decidir cuál es la que más le conviene.
- Un equipo inter-disciplinario facilita el proceso de la generar una solución, esto debido a que cada miembro del equipo abordaría la misma tarea de manera diferente y desde ese punto es más fácil encontrar la solución más eficiente.
- La efectividad de soluciones de big Data se encuentra determinado en mayor medida, en la experticia de los encargados de su implementación que en las herramientas utilizadas.

11.2 Recomendaciones

- Los talleres de Python y API twitter deberían encontrarse en la primera semana del curso, de esta manera se podría coleccionar una mayor cantidad de datos.
- Los alumnos deberían tener acceso a las credenciales de administración de los equipos, con la finalidad de facilitar la realización de configuraciones.

12 Referencias

- [1] P. Alexander, «github.com,» 10 Junio 2016. [En línea]. Available: <https://github.com/alexanderpinchao/BigData/blob/master/twe1.py>.
- [2] P. Alexander, «www.github.com,» Junio 10 2016. [En línea]. Available: <https://github.com/alexanderpinchao/BigData/blob/master/comandosCURL.txt>.
- [3] P. Alexander, «www.github.com,» 10 Junio 2016. [En línea]. Available: <https://github.com/alexanderpinchao/BigData/blob/master/expresiones%20regulares>.

- [4] P. Alexander, «[www.github.com,](https://github.com/alexanderpinchao/BigData/edit/master/clear-parenthesis.py)» 10 Junio 2016. [En línea]. Available: <https://github.com/alexanderpinchao/BigData/edit/master/clear-parenthesis.py>.
- [5] P. Alexander, «[www.github.com,](https://github.com/alexanderpinchao/BigData/blob/master/unificarCoordenadas.py)» 10 Junio 2016. [En línea]. Available: <https://github.com/alexanderpinchao/BigData/blob/master/unificarCoordenadas.py>.
- [6] P. Alexander, «[www.github.com,](https://github.com/alexanderpinchao/BigData/blob/master/cargaBase.py)» 11 Junio 2016. [En línea]. Available: <https://github.com/alexanderpinchao/BigData/blob/master/cargaBase.py>.
- [7] P. Boris, «[www.github.com,](https://github.com/alexanderpinchao/BigData/blob/master/proyecto_big_data)» 10 Junio 2016. [En línea]. Available: https://github.com/alexanderpinchao/BigData/blob/master/proyecto_big_data.
- [8] P. Alexander, «[www.github.com,](https://github.com/alexanderpinchao/BigData/blob/master/credenciales.txt)» 10 Junio 2016. [En línea]. Available: <https://github.com/alexanderpinchao/BigData/blob/master/credenciales.txt>.
- [9] UC, «[www7.uc.c,](http://www7.uc.cl/sw_educ/enferm/ciclo/html/joven/desarrollo.htm)» Universidad de Colombia , 1 Agosto 2008. [En línea]. Available: http://www7.uc.cl/sw_educ/enferm/ciclo/html/joven/desarrollo.htm. [Último acceso: 11 Junio 2016].
- [10] D. m. d. Quito, «[www.quitoteconecta.gob.ec,](http://www.quitoteconecta.gob.ec/faq/)» quitoteconecta, [En línea]. Available: <http://www.quitoteconecta.gob.ec/faq/>. [Último acceso: 11 Junio 2016].
- [11] P. A. Palazzil, «El habeas data en el derecho constitucional,» de *LA DEFENSA DE LA INTIMIDAD Y DE LOS DATOS PERSONALES A TRAVÉS DEL HÁBEAS DATA*, Quito, 2001.
- [12] wikipedia, «[es.wikipedia.org,](https://es.wikipedia.org/wiki/Doble_ciego#Ensayos_a_doble_ciego)» wikipedia, 10 Junio 2016. [En línea]. Available: https://es.wikipedia.org/wiki/Doble_ciego#Ensayos_a_doble_ciego.

13 Anexos

Los scripts generados para el desarrollo de este proyecto se encuentran en el siguiente repositorio de Github:

<https://github.com/alexanderpinchao/BigData>