HARVESTING MYSTERYLAND

Katherine Lasluisa

(Katherine.lasluisa@epn.edu.ec)

Kerly Naranjo

([kerly.naranjo@epn.edu.ec](mailto:kerly.naranjo@epn.edu.ec))

Ricardo Sinchiguano

(Ricardo.sinchiguano@epn.edu.ec)

Escuela Politécnica Nacional

**RESUMEN:**

En el almacén de datos unificado se almacenara los datos recolectados de los diferentes países los datos tendrá diferentes origen (diferentes ordenadores), por el cual la replicación de las bases será necesario para poder obtener una base unificada que nos permita generar informes.

Se utilizara una base de datos no relacional o no SQL, para almacenar la gran cantidad de datos a recolectar, la cual también brinda muchas características para el análisis y procesamiento de toda la información.

La recolección de los datos se realizara sobre el evento “Mysteryland”, festival de música urbana en Holanda, con la cifra de asistentes de 60000 personas.

# INTRODUCCIÓN

En el presente informe se establecerá los puntos y características que corresponden a la primera etapa del proyecto final del semestre en la materia Bases de datos multidimensionales.

En la primera etapa del proyecto final titulado “preparación y exploración de datos” se establece parámetros a tratar, acorde a esto en este documento se desarrollara los temas tales como: definición del propósito sobre el tema escogido para la elaboración del proyecto y el diseño arquitectónico de la fuente de información.

En la definición del propósito sobre el tema escogido, se planteara puntos de vistas diferentes de cada uno de los integrantes del grupo de trabajo a cargo de este proyecto, con una lógica orientada a obtener conocimientos en “Data WareHouse” y los temas relacionados dentro de este.

En el diseño arquitectónico de la fuente de información, se estudiara las fuentes de información con una arquitectura estructurada, no estructurada y semiestructurada.

Los datos recolectados se almacenaran creando consigo un almacén de datos, en el cual nos permitirá acceder a los datos, además de realizar consultas multidimensionales utilizando el procesamiento analítico en línea (OLAP) el cual nos permite agilizar consultas en grandes cantidades de datos (Big Data).

Los datos recolectados deberán pasar por un proceso de depuración en el cual se asegura que cuenten con características tales como: consistencia, disponibilidad y trazabilidad.

# Objetivo General

* Generar una recopilación de tweets analizando los resultados en diferentes países de América Latina.

**Objetivo Especifico**

# Recopilar tweets acerca de Mysterland, para agruparlos por países y obtener resultados de cada uno de ellos.

* Demostrar los resultados obtenidos mediantes gráficos estadísticos para representar los datos de manera visual.

# Marco Teórico

* Sistema de soporte de decisiones (DSS).- es un conjunto de técnicas y herramientas para tratar y analizar datos y para apoyo a la toma de decisión. [1]
* Big Data.- es el conjunto de datos grandes que aplicaciones informáticas tradicionales de procesamiento de datos no son capaces de procesar utilizando método tradicional para encontrar patrones repetitivos dentro del mismo. [3]
* Almacén de datos.- es una colección de datos que soporta los procesos de toma de decisiones, con características tales como: orientado a objetos, integridad y coherencia, no volátil. [1]
* Procesamiento analítico en línea (OLAP).- es una tecnología que se usa para organizar grandes bases de datos empresariales y admitir la inteligencia empresarial. Las bases de datos OLAP se dividen en uno o más cubos. Cada cubo lo organiza y diseña un administrador de cubos para que se adapte a la forma en la que recupera y analiza los datos con el fin de que sea más fácil crear y usar los informes de las tablas dinámicas y los gráficos dinámicos. [2]
* Meta datos.- son los datos utilizados para definir otros datos, en el cual se especifica la fuente, valores, uso y características de los datos del almacén de datos y define como los datos se pueden cambiar y procesar en cada capa de la arquitectura. [1]

# Implementación

* **Propósito del tema**

Lollapalooza, Coachella, Tommorrowland son algunos de los eventos más importantes de música electrónica del mundo, se realizan durante todo el año en diferentes países. Mysteryland es uno de estos grandes eventos, a realizarse en el mes de agosto, hasta el momento se estima la asistencia de 6000 aficionados.

La arquitectura propuesta para el diseño de la base de datos NO SQL, está basado en la recolección de tweets en couchDB

* **Arquitectura de la base de datos**

La recolección de tweets se realizó en diferentes países, como por ejemplo: Ecuador, Estados Unidos y Holanda. Ya que por tal motivo la recopilación se realizó individualmente de acuerdo a cada uno de los países antes mencionados.

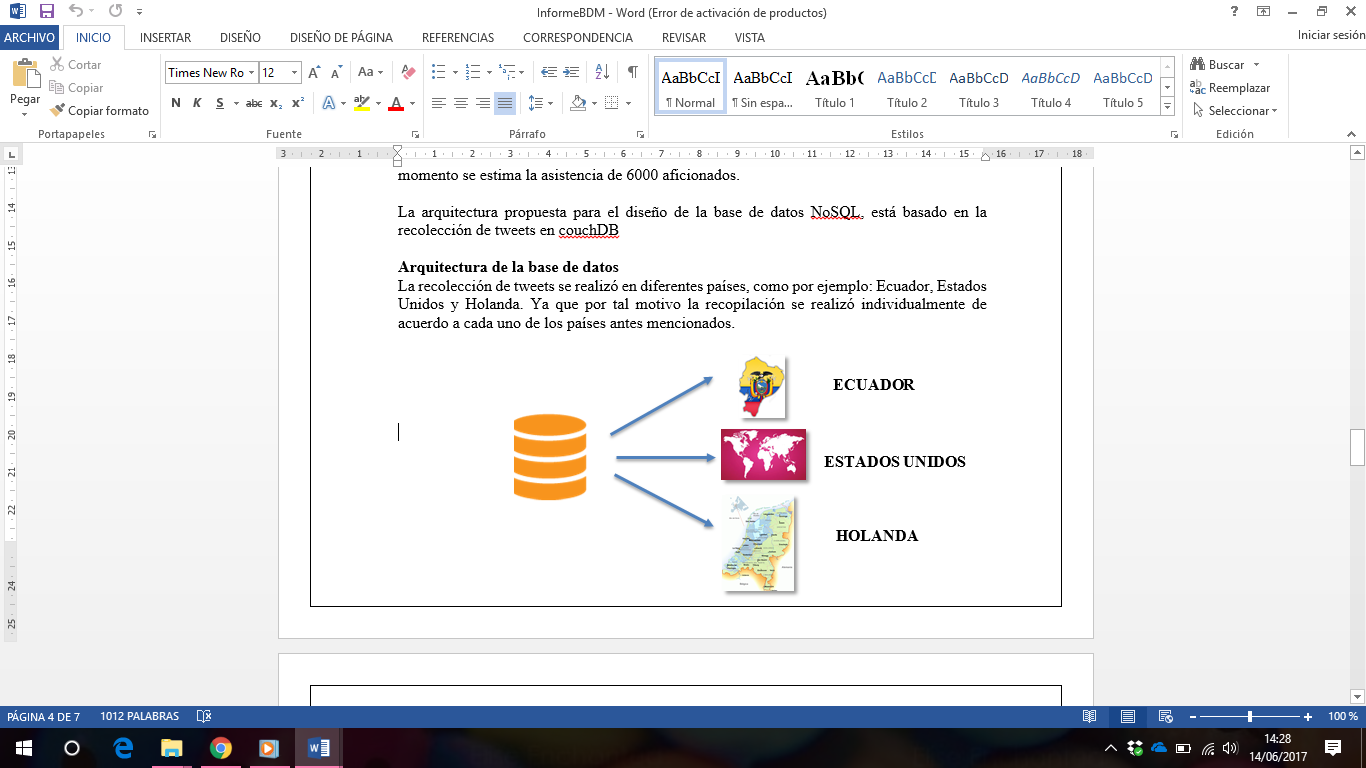


Figura 1. Arquitectura de la base.

Harvesting: El harvesting o cosecha de Tweets es una actividad que no se la puede realizar manualmente, po

* **Application Settings and Acces Token**

Contiene las siguientes credenciales:

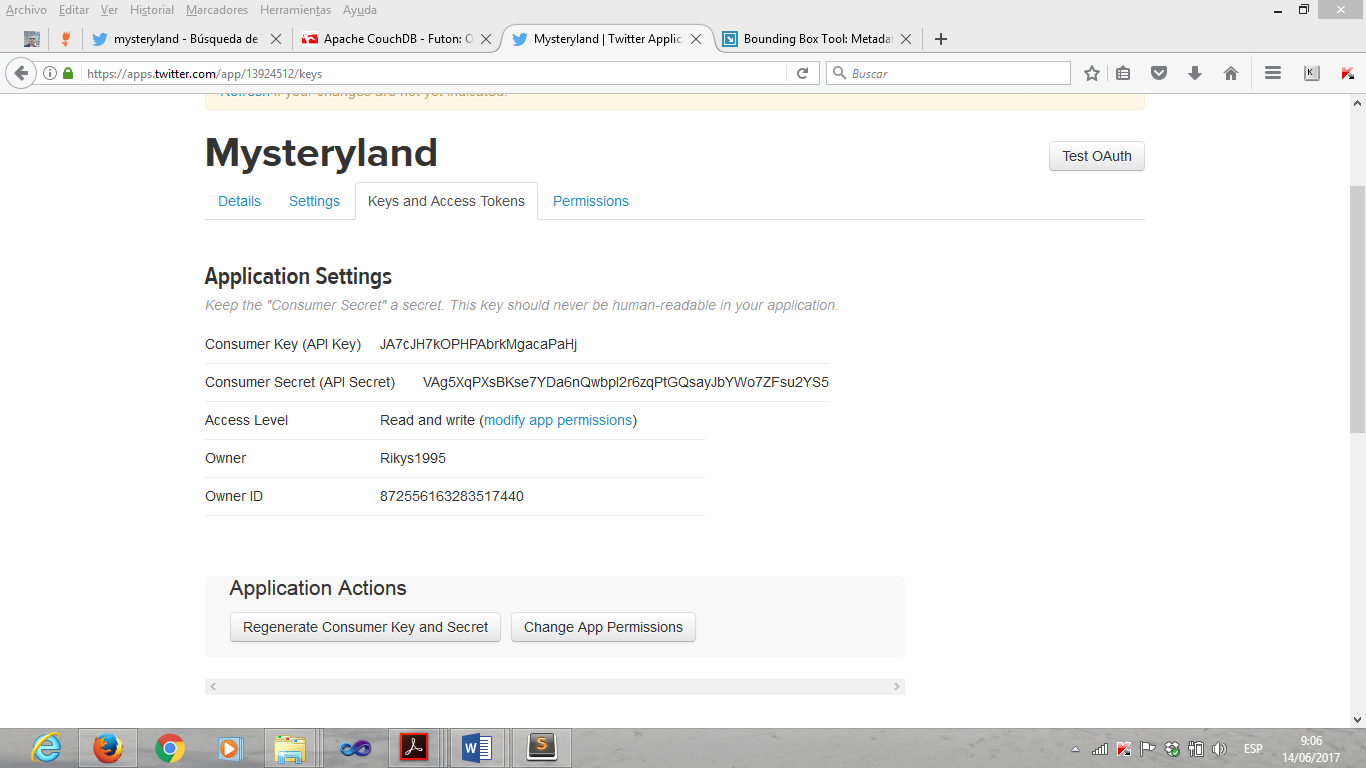


Figura 2. Credenciales Consumer.

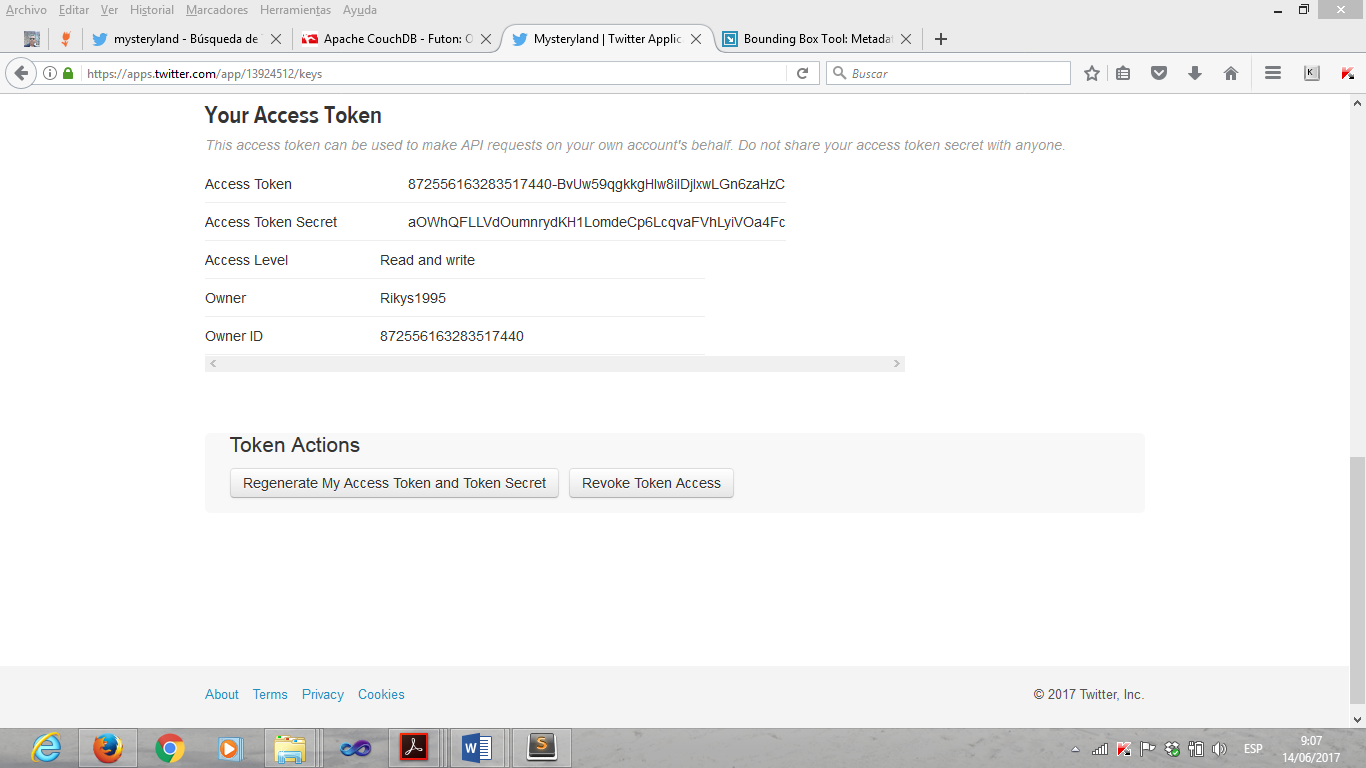


Figura 3. Credenciales Acces.

* **Bouxing Box**

Debemos acceder al siguiente enlace <http://boundingbox.klokantech.com/>

El formato en el cual debemos capturar la localización es CSV

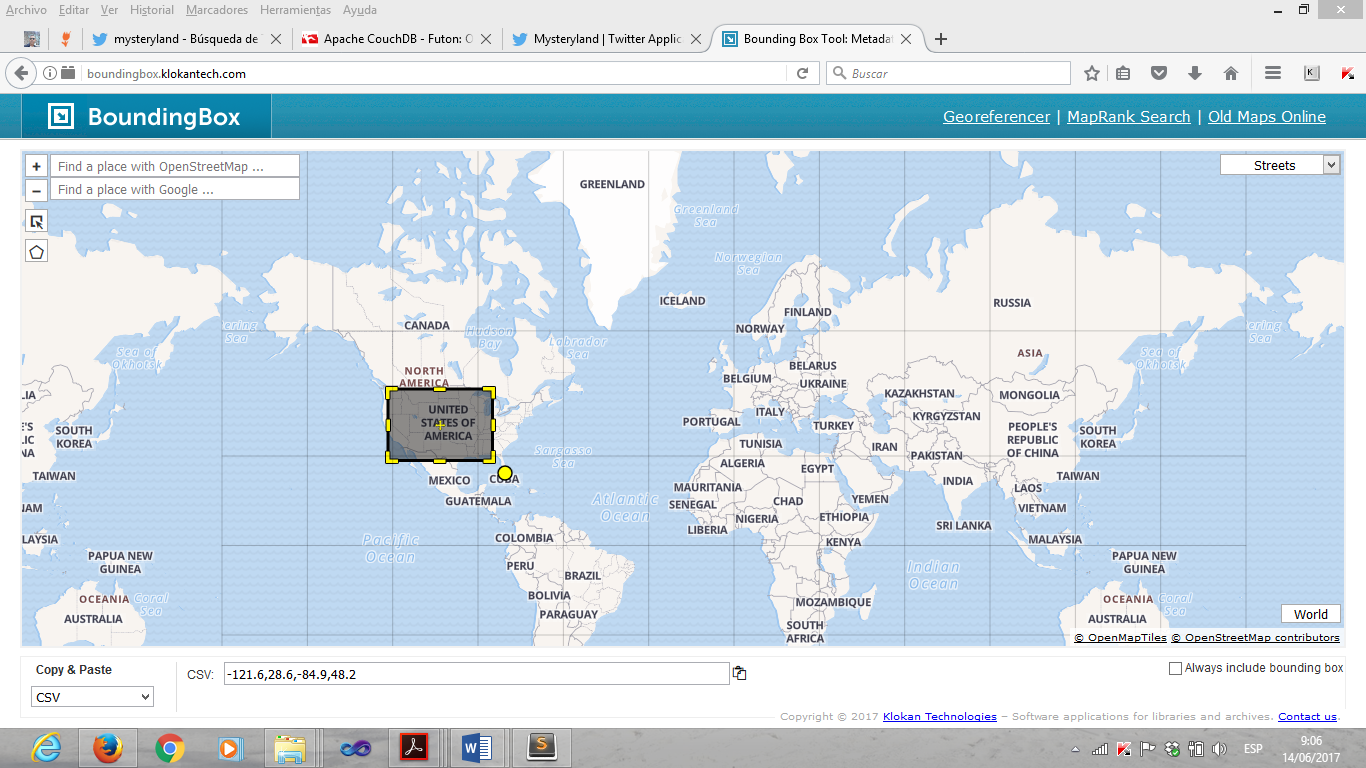


Figura 4. Bouxing Box

* **Recolección de tweets**

En el archivo: tweets.py, editamos las credenciales que se obtuvo el momento de crear la Application:

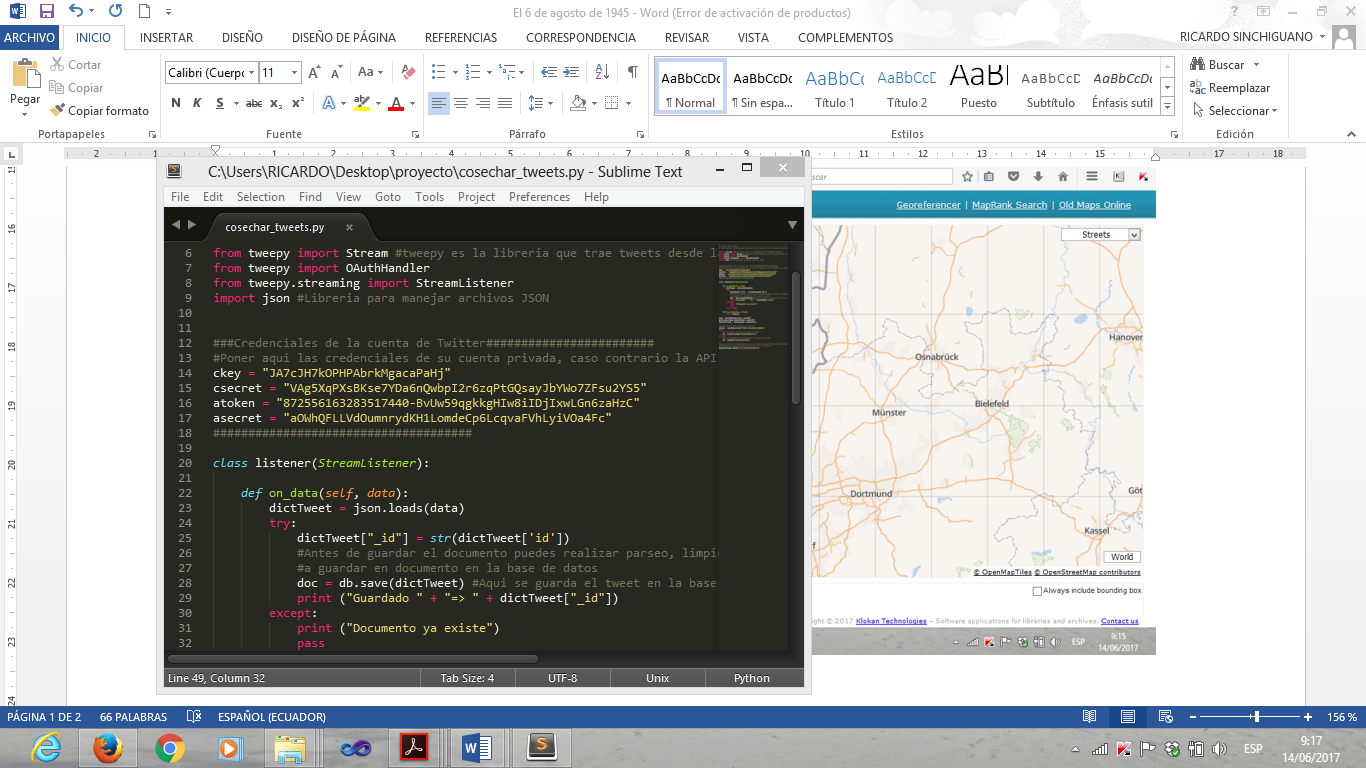


Figura 5. Recolección de tweets

Conexión con nuestro servidor CouchDB y creación de la base de datos

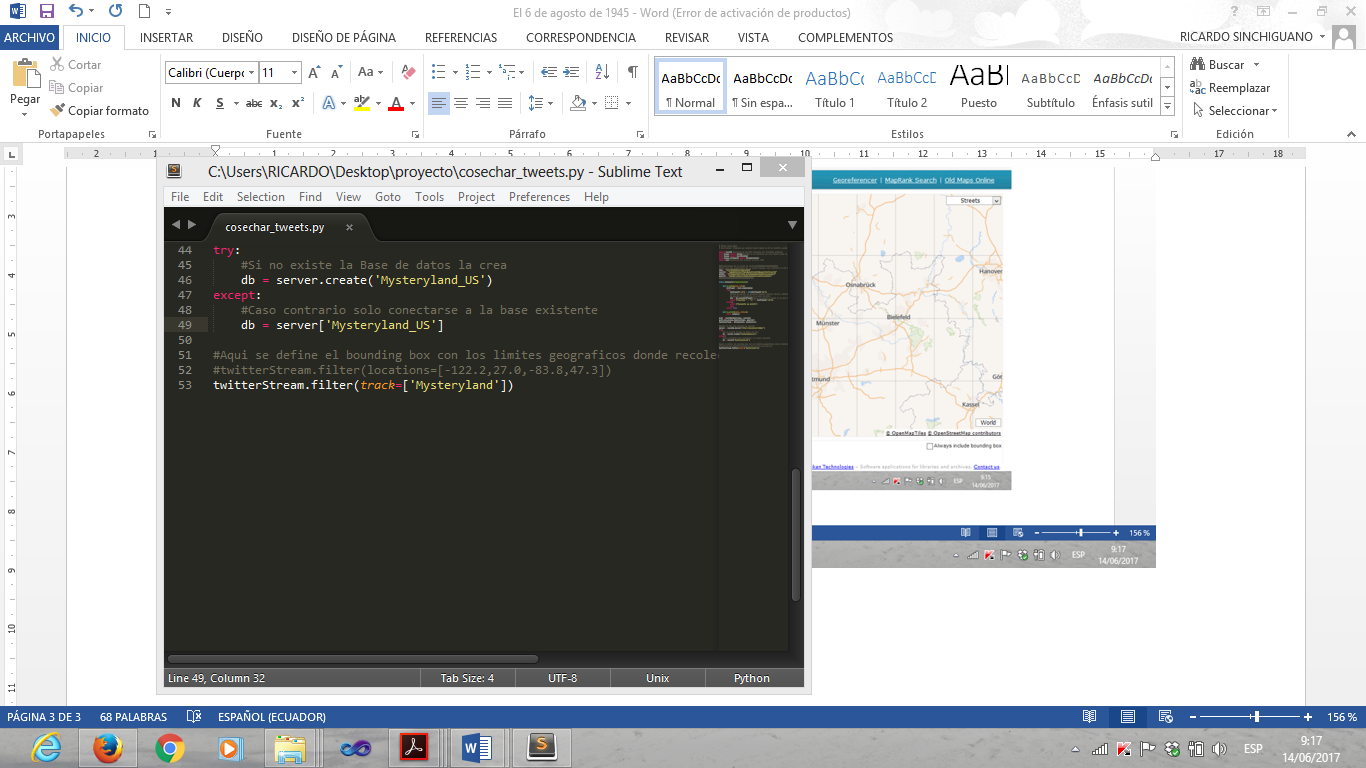


Figura 6. Conexión.

Resultados de la recolección de datos (2 horas)

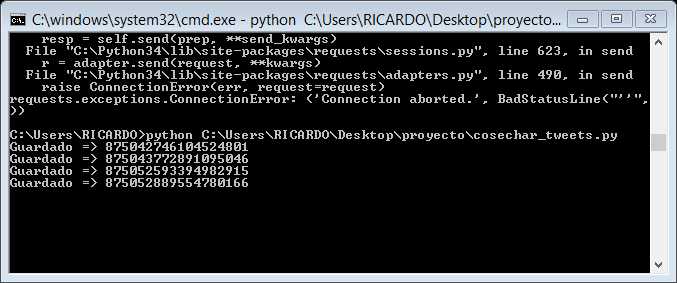


Figura 7. Resultados de recolección.

Almacenamiento de los datos en CouchDB

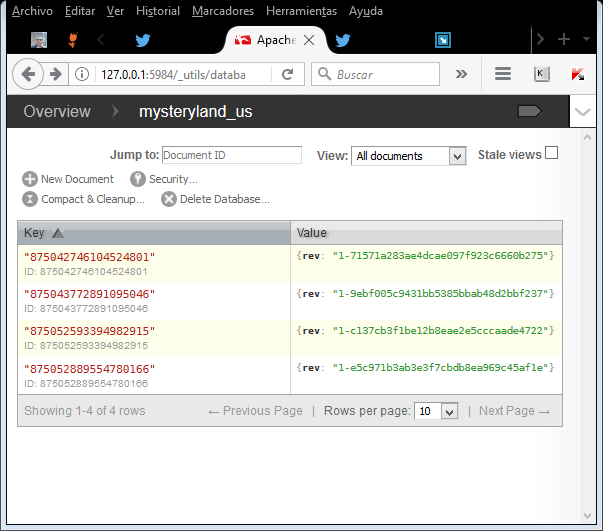


Figura 8. Base primaria.

# CONCLUSIONES

* La recopilación de datos a través de una API de twitter nos permitió construir el proceso de data warehouse obteniendo los resultados de tweets esperados. Hasta el momento el evento Mysteryland cuenta con un gran número de tweets, se espera cuadriplicar este número hasta un mes antes del evento.
* El total de información recopilada permite estimar el número de jóvenes y adolescentes de diferentes partes del mundo que están interesados en el evento Mysteryland a realizarse el mes de agosto del presente año. Al darle seguimiento a dicha información podremos estimar el número de asistentes, la popularidad del evento, entre otros datos de interés para organizadores o personas involucradas.

# REFERENCIAS

[1]2017. [Online]. Available: https://dev.twitter.com/rest/tools/console. [Accessed: 14- Jun- 2017].

[2]2017. [Online]. Available: http://127.0.0.1:5984/\_utils/index.html. [Accessed: 14- Jun- 2017].

[3]"Bounding Box Tool: Metadata Enrichment for Catalogue Records by Visually Selecting Geographic Coordinates (Latitude / Longitude) for Maps", Boundingbox.klokantech.com, 2017. [Online]. Available: http://boundingbox.klokantech.com/. [Accessed: 14- Jun- 2017].