**EXTRACCIÓN Y OBTENCIÓN DE DATOS**

**IVONNE VIRGINIA VEGA PEDRAZA**

Ivonne.vega01@epn.edu.ec

**JOSE GEOVANNY DIAS AMAGUAÑA**

Jose.dias@epn.edu.ec

**ESCUELA POLITÉCNICA NACIONAL**

**ESCUELA DE FORMACIÓN DE TECNÓLOGOS**

*RESUMEN*— En el presente documento se mostrará el diseño de un Data Warehouse en el cual convergen diferentes fuentes de información como Tableau, Twitter y Data.gov

1. INTRODUCCIÓN

U

n data warehouse es un repositorio unificado para todos los datos que recogen los diversos sistemas de una empresa. El repositorio puede ser físico o lógico y hace hincapié en la captura de datos de diversas fuentes sobre todo para fines analíticos y de acceso. [1]

Tableau es la plataforma de análisis integral más eficaz, segura y flexible para los datos. Además es la única plataforma de inteligencia de negocios que convierte los datos en información útil. [2]

Twitter es un servicio de comunicación bidireccional con el que puedes compartir información de diverso tipo de una forma rápida, sencilla y gratuita. [3]

Data.gov es un sitio web del gobierno de los EE.UU, apunta a mejorar el acceso público a conjuntos de datos de alto valor, legibles por máquina, generados por el Poder Ejecutivo del Gobierno Federal. El sitio es un depósito de información del gobierno federal, estatal, local y tribal, disponible para el público. [3]

1. OBJETIVOS

* Obtener información de tres fuentes distintas.
* Unificar la información de las diferentes fuentes.
* Visualizar en una interfaz web la información obtenida.

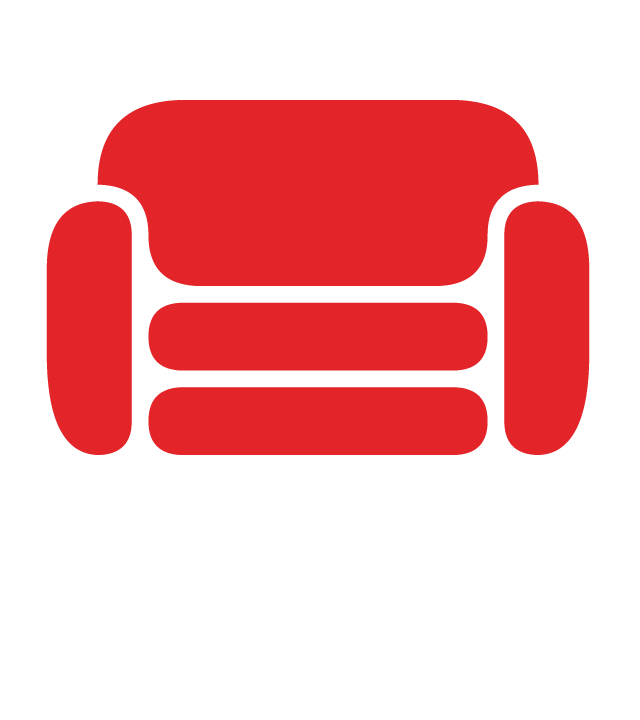
1. DESARROLLO

Se han seleccionado 3 temas para la recolección de datos:

1. Mascotas.
2. Entretenimiento.
3. Tecnología.

Para realizar este proyecto se utilizó los siguientes recursos.

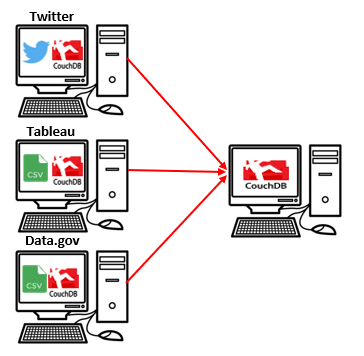
1. CouchDB 2.2.
2. Phython 5.0
3. USBWebserver v8.6 (Portable).
4. Editores de texto (Brackets).
5. Navegador Web

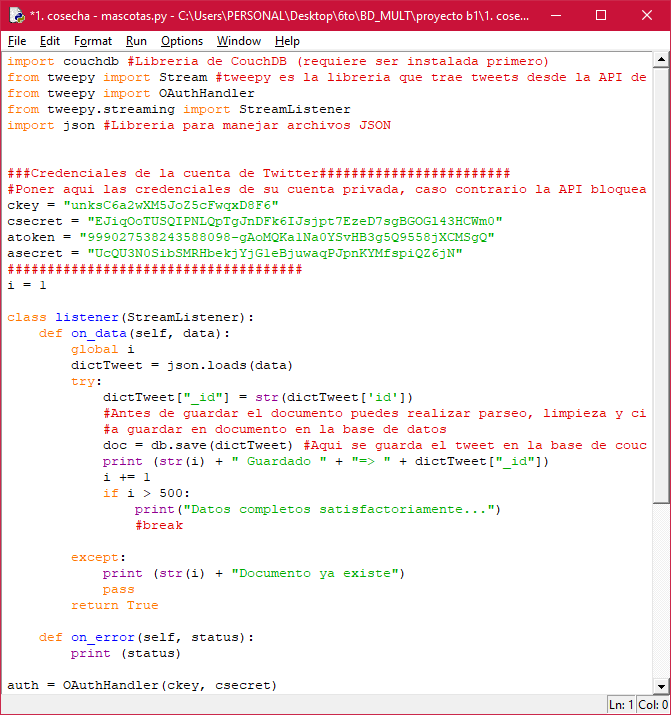


1. ARQUITECTURA



1. Recolección de información de Twitter.

Para la recolección de datos de Twitter se utilizó el siguiente script:



En el cual tenemos que setear la URL del servidor de couchDB:



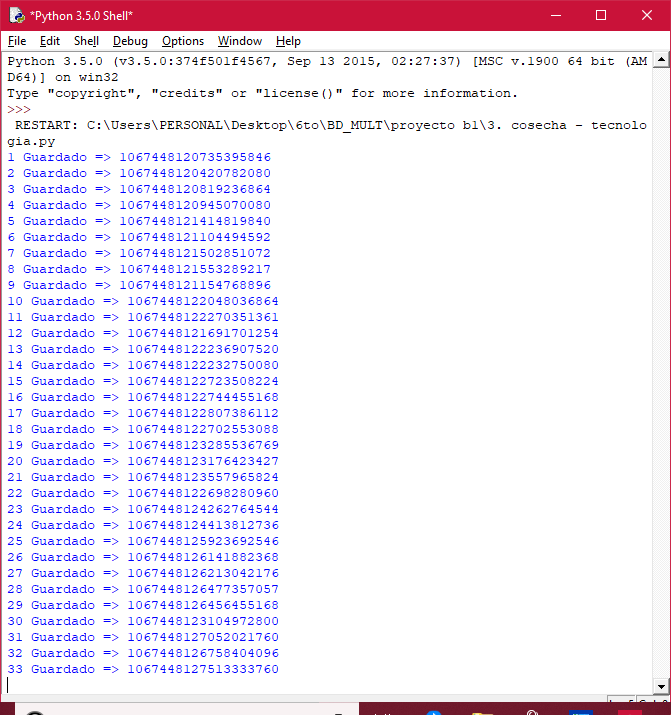
Crear la base de datos:

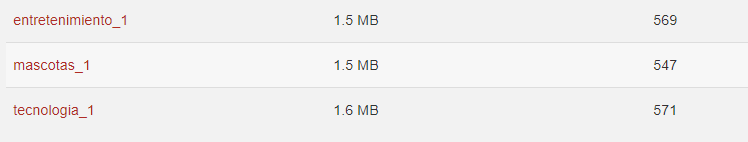


Definir el filtro por el cual se quiere buscar los tweets:



Al momento de ejecutar los scripts se pudo observar como se iba guardando los datos en couchDB.





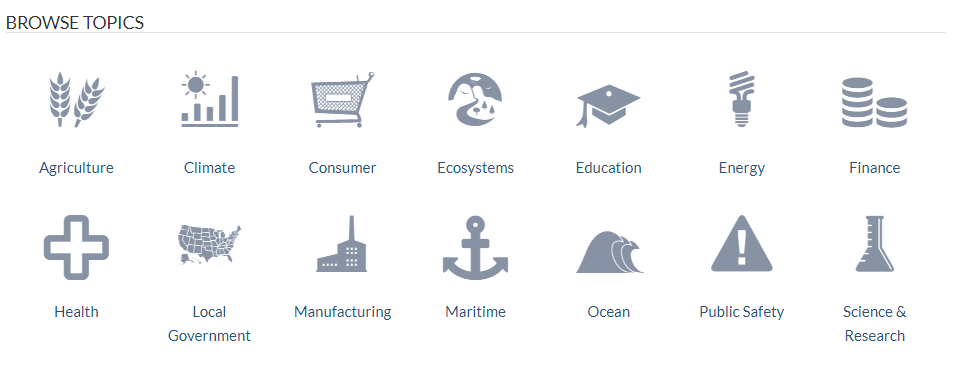
1. Recolección de datos de Tableau.

Para obtener los datos de Tableau se ingreso a la siguiente página: <https://www.tableau.com/es-es> , en la cual se procedió a descargar los archivos Excel.

1. Recolección de datos en Data.gov

Para obtener los datos de Data.gov se ingresó a la siguiente página:

<https://www.data.gov/> en la cual se puede escoger datos entre diversos temas, dichos documentos pueden ser archivos CSV o Excel.

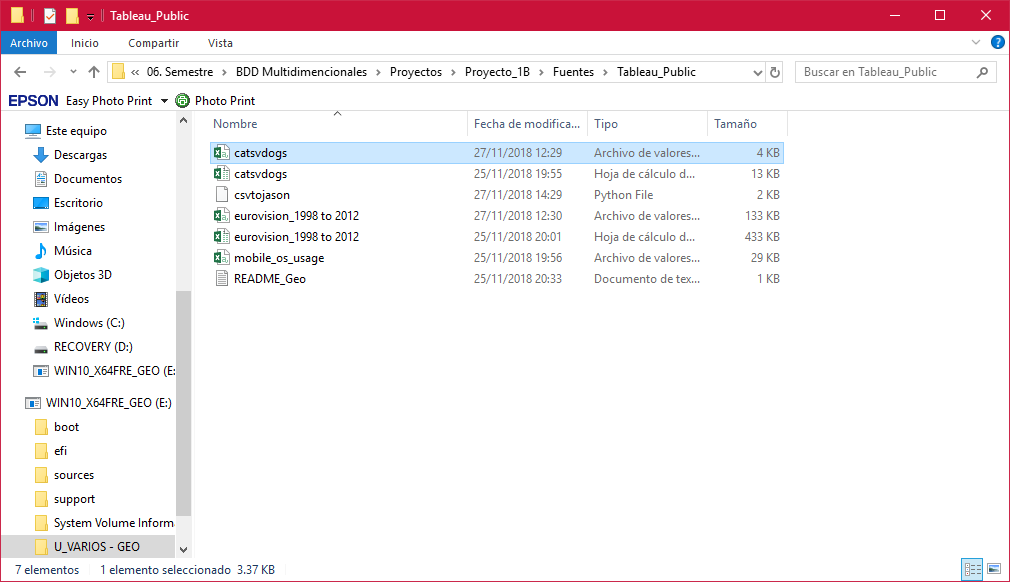


1. Transformar archivos Excel a CSV y guardar en CouchDB con formato JSON.

Primero tenemos que transformar los archivos de Excel en CSV. Simplemente al archivo \*.xlsx lo guardamos como CSV (delimitados por comas).

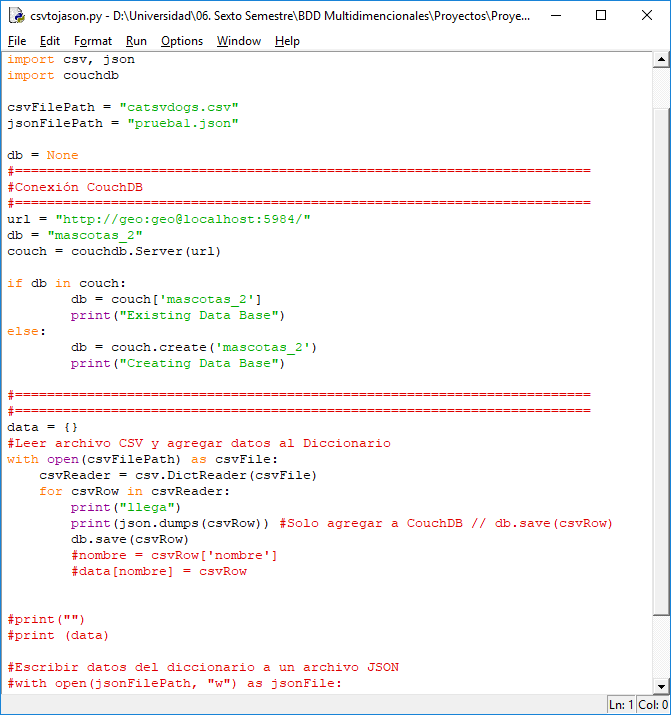


Se nos guarda el archivo en \*.csv con el cual podremos ingresar a nuestra base de datos NoSQL.



1. Ingreso de datos de archivo CSV a CouchDB.

Para realizar este proceso hemos escrito un pequeño Script en lenguaje Python.



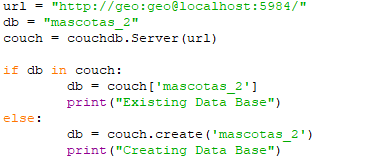
En el cual importamos las librerías necesarias para poder ejecutar el código y no nos marque error.



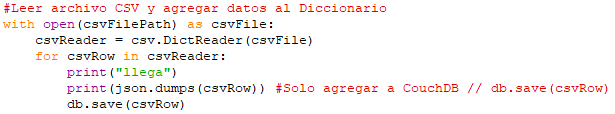
Se ingresa el nombre del archivo con la extensión \*.csv, el cual contiene el dataset que se guardará en la base de datos.



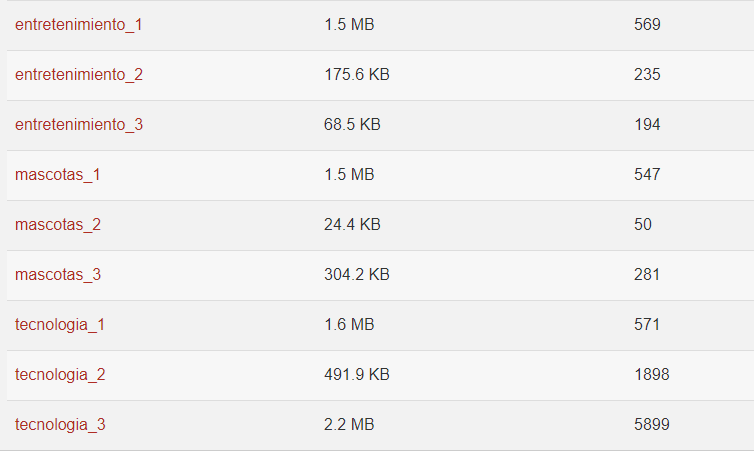
Realizamos la conexión a CouchDB y creamos la base de datos con el nombre seleccionado.



El Script leerá cada línea del documento CSV en formato JSON e irá guardando automáticamente en CouchDB mediante un ciclo “For”.

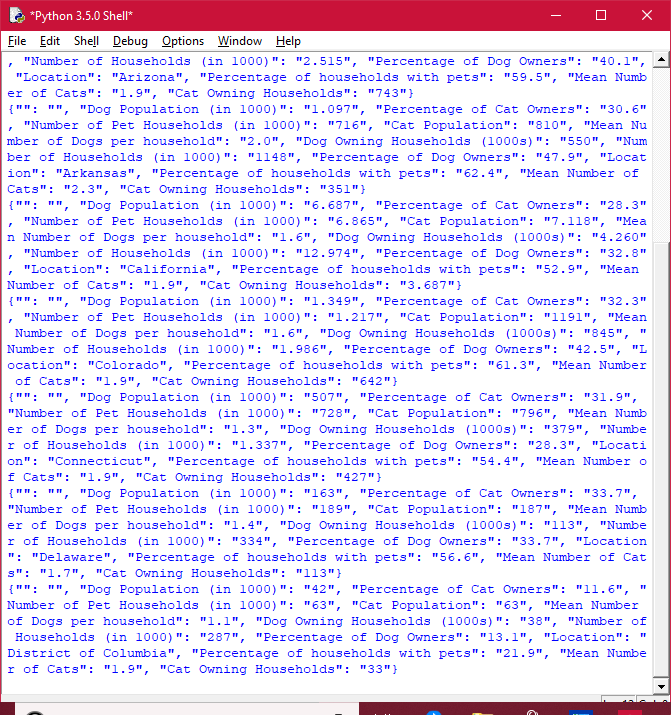


Ahora podemos observar cómo se crea la base de datos y como se van guardando los datos:



**Fuente: Tableau**

Tema: Mascotas



Tema: Entretenimiento



Tema: Tecnología



**Fuente: Data.gov**

Tema: Mascotas



Tema: Entretenimiento



Tema: Tecnología



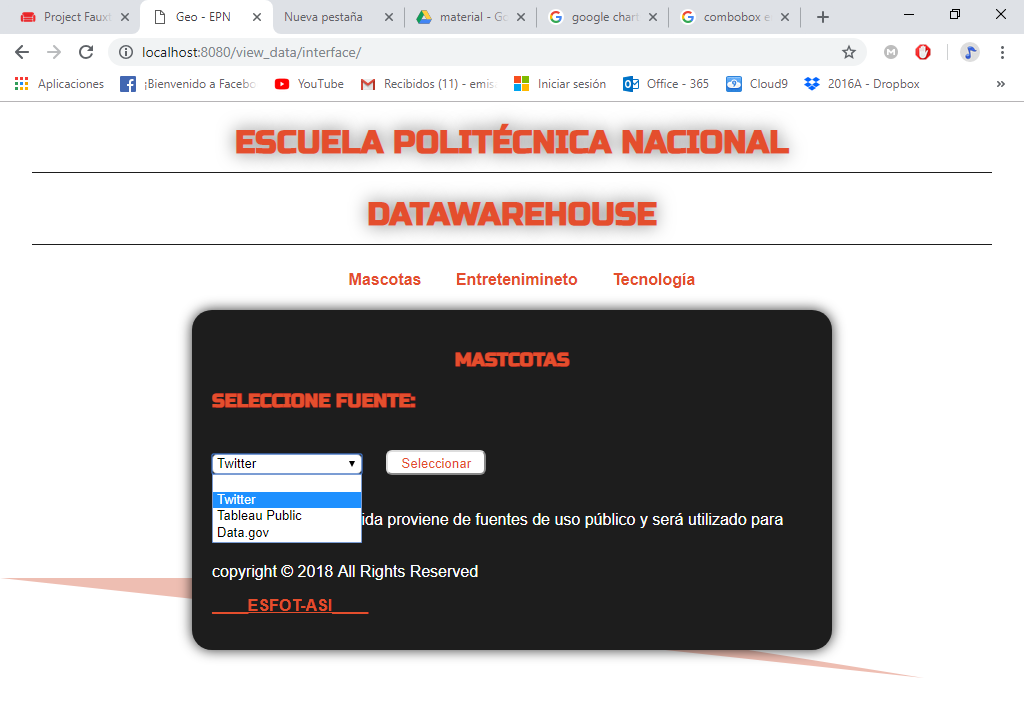
Cada vez que ejecutábamos un Script se visualizaba como los datos se leían en formato JSON y se guardaban en la base de datos NoSQL de CouchDB.

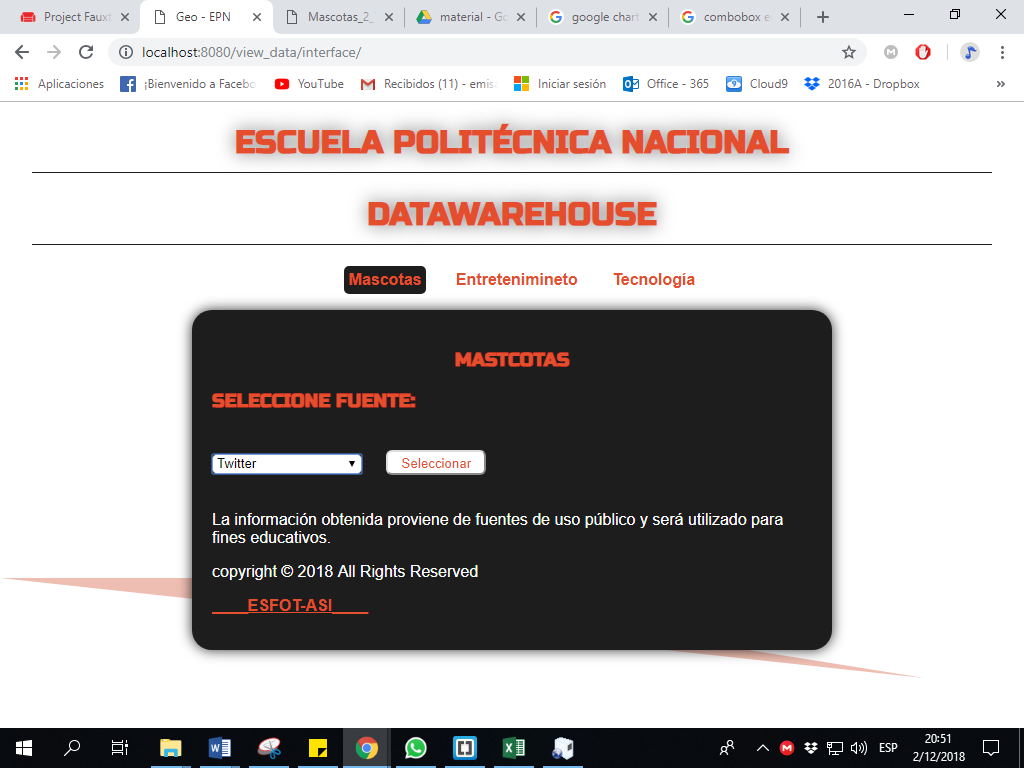
1. APLICACIÓN WEB.

Se creó una interface interactiva para poder visualizar la información recopilada de las diferentes fuentes.

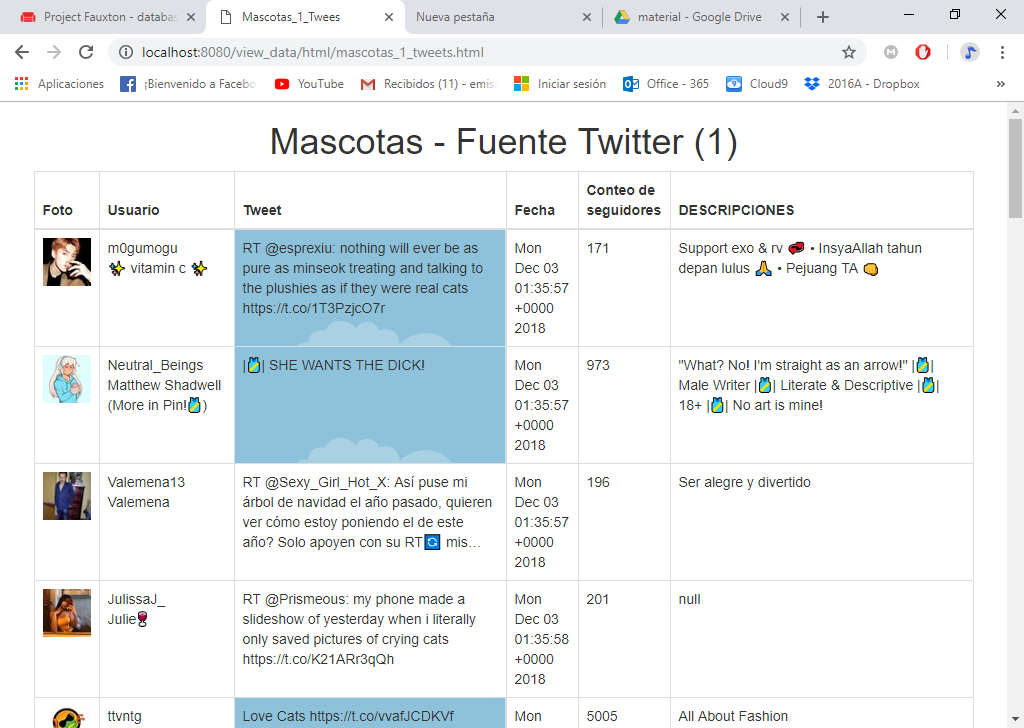
La interface cuanta con un menú de navegación que permitirá al usuario navegar de forma fácil y segura.

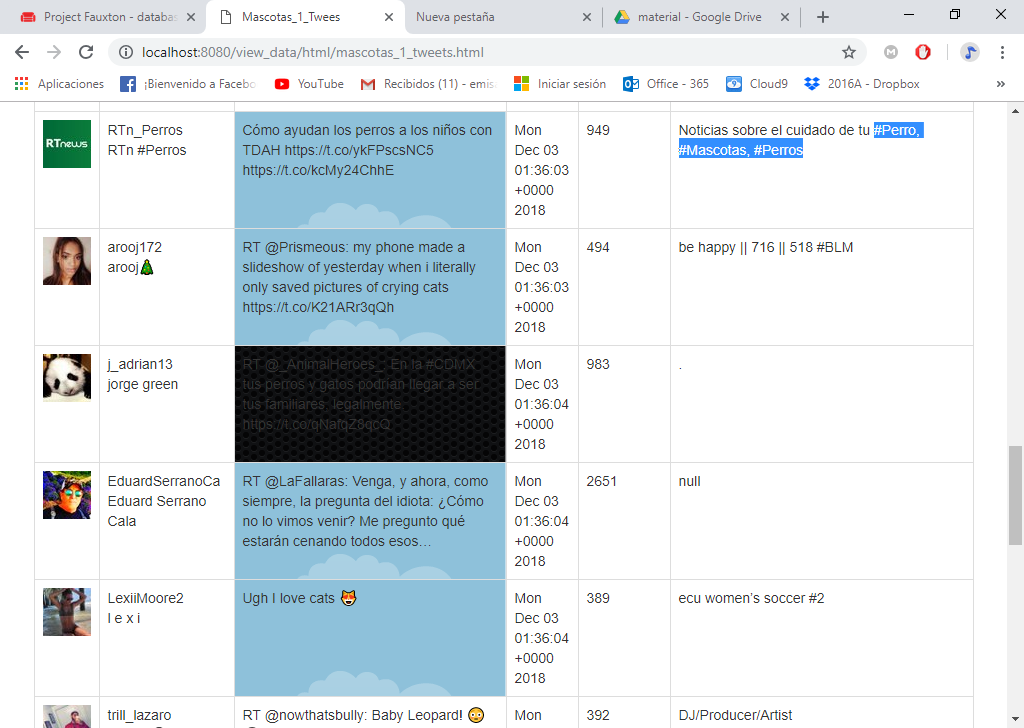
Se puede seleccionar el tema de interés y mediante un selector de etiquetas elegir la fuente de donde fue obtenida la información.





Seleccionamos la fuente y la interfaz nos direccionara a una nueva página donde mostrara la información de guardada en nuestra base de datos NoSQL de CouchDB.





1. CONCLUSIONES

Para diseñar una buena arquitectura de un Data Warehouse es necesario conocer bien los requerimientos del negocio y hacer un estudio profundo de las fuentes externas que nos van a suministrar los datos.

Tener un Data Warehouse dentro de una organización es muy útil debido a que ayuda y facilita la toma de decisiones ya que permite tener un buen control sobre toda la información que necesita la organización para poder tomar decisiones, basándose en datos confiables y que pueden tener a la mano en cualquier momento.

1. REFERENCIAS

|  |  |
| --- | --- |
| [1] | «PowerData,» PowerData, [En línea]. Available: https://www.powerdata.es/data-warehouse. [Último acceso: 29 11 2018]. |
| [2] | «Tableau,» Tableau, [En línea]. Available: //www.tableau.com/es-es/products/what-is-tableau#UWd33zZDTkMuUIYY.99. [Último acceso: 29 11 2018]. |
| [3] | «WebEmpresa,» [En línea]. Available: https://www.webempresa.com/blog/que-es-twitter-como-funciona.html. [Último acceso: 29 11 2018]. |
| [4] | «Wikipedia,» [En línea]. Available: https://en.wikipedia.org/wiki/Data.gov. [Último acceso: 29 11 2018]. |
| [5] | M. Rouse, «SearchDataCenter,» [En línea]. Available: https://searchdatacenter.techtarget.com/es/definicion/CouchDB. [Último acceso: 03 11 2018]. |

1. ANEXO

<https://github.com/GeovannyDias/Datawarehouse-with-CouchDB>