Proyecto Final

Bolaños Erick, Heredia Alfonso, Pizarra Jhonathan
Quito – Ecuador
Escuela Politécnica Nacional
Base de Datos Multidimensionales
erick.bolanos@epn.edu.ec, alfonso.heredia@epn.edu.ec, jhonathan.pizarra@epn.edu.ec

I. INSTRUCCIONES:

Realizar un caso de estudio para las siguientes temáticas:

- Tráfico vehicular en las 5 principales ciudades del Ecuador.
- Eventos deportivos en los principales estadios de Ecuador.
- Pulso político en 5 ciudades de Ecuador.
- Top 10 twitteros en 5 ciudades de Ecuador.
- Top 10 quejas en el Ecuador.
- Actividades y hobbies.
- Conciertos y eventos públicos.
- Tema definido por el estudiante.
- Restaurantes y sitios de esparcimiento.
- Eventos o noticias mundiales.

La recopilación de información se puede realizar con geolocalización o con filtro de palabras.

Si se utiliza geolocalización se recomienda subdividir la región. Si se utiliza filtro de palabras se recomienda usar varias palabras por script.

Se debe crear la indexación en al menos 2 nodos. (debe crear un clúster)

Para la visualización se creará una página web en la cual se insertará un dashboard (puede ser de kibana o Google charts)

Cada caso de estudio debe tener su propio dashboard y de todo el proyecto al menos una visualización debe tener geolocalización.

II. DEFINICIÓN DEL CASO DE ESTUDIO

1. Para el caso de Tráfico vehicular en las cinco principales ciudades se planteó la extracción de datos analizando las principales ciudades como son Quito, Guayaquil, Cuenca, Ambato y Loja con la ayuda de las páginas que estén relacionadas con las agencias de transito de dichas ciudades.

2. Al analizar el caso que consiste en eventos deportivos de los diferentes estadios es necesario tener presente los estadios en los cuales estos se van a presentar, así como el rango geográfico en el cual se puede obtener información de dichos eventos atreves de twitter o a su ves buscar una página en la cual dispongan de dicha información e introducir los datos manualmente en una base de Datos.

3. Para analizar los pulsos políticos en el Ecuador se tomó en cuenta los que creemos más representativos o son más conocidos ya sea por su participación, apoyo o popularidad de los cuales hemos tomado los siguientes, Juntos Podemos, Partido Social Cristiano, Pachakutik, Alianza PAIS, Suma, entre otros menos conocidos.

4. Para tener un mejor hallazgo de datos del top de 10 twitteros, seleccionamos las siguientes palabras clave que se ejecutaran en el script: top, Ecuador, Quito, Guayaquil, Cuenca, Loja y Ambato. Siendo así, se empleo una API de Twitter que migraba los datos a CouchDB a través de un script de Python.

5. En el caso de las quejas, recolectamos datos de fuentes estructuradas como el INEC, Datos abiertos, entre otros. Sin embargo, la información fue poca, y optamos por recopilar estos datos con bases NOSQL, a través de una cosecha de tweets. La cosecha si hizo hacia MongoDB.

6. Actividades y hobbies fue de los más fáciles de conseguir, eso también desencadena en un problema puesto que, al existir tantas fuentes de datos diferentes, no siempre suelen ser “verídicas”. En todo caso, lo que está a nuestro favor, es que trabajar este tipo de Datassets para bases de datos SQL fue relativamente fácil, puesto que el proceso de extracción y carga se agilizo al tener este Datasset una sola columna.

7. En la parte de conciertos, los CSV mostraban datos como fechas, artistas, y espacios. A simple vista en el CSV observábamos que se iba repitiendo el espacio, o sea, existan mas registros sobre el lugar del concierto y este era el coliseo general Rumiñahui. Por lo que, para las visualizaciones teníamos ya una idea remota de que hubiera pasado si filtrábamos este campo.

8. De este tema es del que se hará una visualización de tipo Geilocalizacion. Es de nuestro interés saber sobre los países que más generan noticias, por lo que empleando una cosecha de tweets recopilaremos información sobre diferentes países, por lo que sucede en las noticias en canales de tv, se observa mucho que las noticias provienen mayormente de Chile, EEUU, y últimamente de China. Con lo que, al hacer el análisis comprobaremos nuestra hipótesis. Esto, involucra el uso de bases NOSQL, y será una base de MongoDB quien contenga estos datos.

9. En esta ocasión se planeta presentar una visualización de los restaurantes y sitios donde la gente mayormente acude a degustar. Éste tipo de información fue uno de los dataset más difíciles de conseguir debido a que la interpretación de restaurantes había en gran cantidad junto con datos no tan precisos, es decir no se parecían entre sus atributos y dificultaba su uso.

10. Sería muy similar al tema número 8, ya que también involucra noticias. En este caso lo que diferencia es por ejemplo la temática, homicidios, accidentes, noticias en si, ya no solo saber qud países. En todo caso se hará manejo de bases Nosql para recopilar esta información. Después se parara a un Clúster para que finalmente llegue a un visualizador, que sería el de Mongo.

El filtro de palabras tendrá claves como Noticias, BBC, News, entre otras que agilizan el proceso de búsqueda. De este tipo de datos, hay muchos, y es por eso que usaremos una base nosql, datos con propiefsdes tales como volumen, y variedad, además de que estos se generan todos los días.

III. OBJETIVOS

a. Objetivo General

- Crear un DataWarehouse

b. Objetivos específicos

- Recopilar información de diferentes fuentes
- Creac un cláuster
- Aplicar el principio de ETL(Extracción, Limpieza y carga) de los datos recopilados
- Analizar factores influyentes en los datamark

IV. DESCRIPCIÓN DEL EQUIPO DE TRABAJO Y ACTIVIDADES REALIZADAS POR CADA UNO

Erick se encargará de la recolección de datos haciendo uso de las bases NOSQL, por otro lado, yo, Jhonathan Pizarra haré la recolección con las bases tradicionales, también conocidas como SQL, por otra parte, Alfonso se encargará de recopilar datos de fuentes internet haciendo uso de Rapidminer.

Todos irémos realizando el informe y explicaremos sobre cada punto de desarrollo. También todos crearemos videos en Youtube para que ayude al veedor en algún trabajo puntual. Sin embargo, lo recomendable sería ver todos los videos de todos los estudiantes para no perder orientación del proyecto.

A mi cargo estarían los Script y Datasset de los incisos 4, 5, y 6 del proyecto.

V. CRONOGRAMA DE ACTIVIDADES

N..	Actividades	Fechas												
		29/01/2020	30/01/2020	31/01/2020	01/02/2020	05/02/2020	06/02/2020	07/02/2020	08/02/2020	09/02/2020	10/02/2020	11/02/2020	12/02/2020
1	Planificacion												
2	Instalacion de heramientas												
3	Analisis de heramientas												
4	Diseño de arquitectura												
5	Busqueda de BD y extraccion de datos												
6	Integracion de BD												
7	Visualizacion y Analisis												
8	Creacion de informe												
9	Edicion de video												
10	Presentacion												

VI. ASIGNACIÓN DE ACTIVIDADES A CADA MIEMBRO DEL EQUIPO

Erick realiza la abstracción de datos de los casos de estudio 4, 9 y 10, empleando bases de datos no relacionales, obteniendo sus datos desde Twitter.

Actividades
Heredia
✓ Instalar elasticsearch, kibana, losgtash, cerebro, Mongo DB, Couch DB. ✓ Recopidar 3 casos de estudio. ✓ Abstraer datos de redes sociales ✓ Establecer objetivos ✓ Configuración en logstash
Bolaños:
✓ Instalar elasticsearch, kibana, losgtash, cerebro, Mongo DB, Couch DB. ✓ Diseñar la arquitectura de solución Data Lake ✓ Recopidar 3 casos de estudio. ✓ Recopilar datos en elasticsearch ✓ Realizar Informe ✓ Recopilar base de datos en Dvd ✓ Detallar Readme
Pizarra:
✓ Instalar elasticsearch, kibana, losgtash, cerebro, Mongo DB, Couch DB. ✓ Recopilar 4 casos de estudio. ✓ Abstraer datos de redes sociales. ✓ Establecer objetivos ✓ Definir los desafíos y problemas encontrados ✓ Grabar y editar video

VII. RECURSO Y HERRAMIENTAS UTILIZADAS

Bases de datos SQL

- MySQL
- SQL server

Bases de datos NoSQL

- MongoDB
- CouchDB

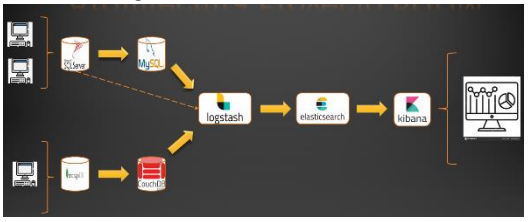
Fuentes de internet

- Twitter
- Webscraping

Concentrador de datos

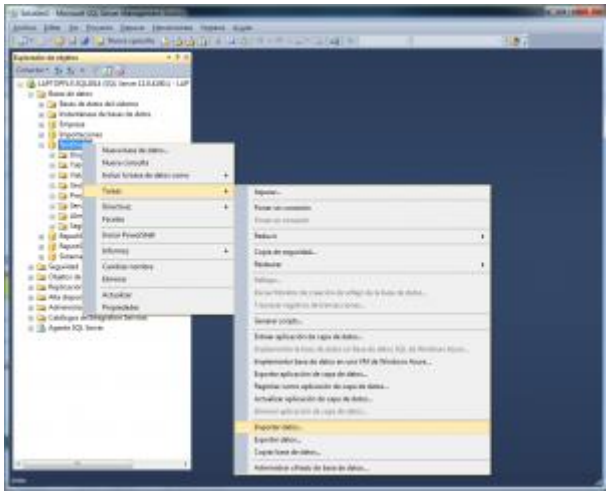
- Elasticsearch

VIII. ARQUITECTURA DE LA SOLUCIÓN



IX. EXTRACCIÓN DE DATOS

En SQL Server Management Studio seleccionar la base de datos donde se importarán los datos, pulsar botón derecho, seleccionar Tareas (Task) y luego Importar datos (Import Data)

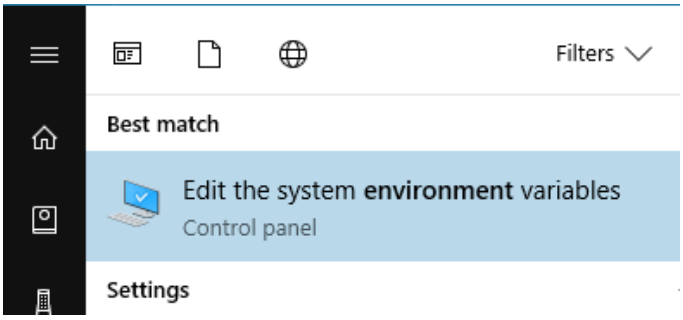


2. Aparece la ventana de inicio del asistente de importación de datos. Pulsar Siguiente

3. Seleccionar el origen de datos, en este caso se seleccionará Microsoft Excel, luego seleccionar el archivo del disco donde se encuentran los datos. Note que se encuentra activada la casilla de verificación «La primera fila tiene nombres de columna» lo que va a definir los nombres de campo en la tabla al finalizar la importación. Pulsar Siguiente.

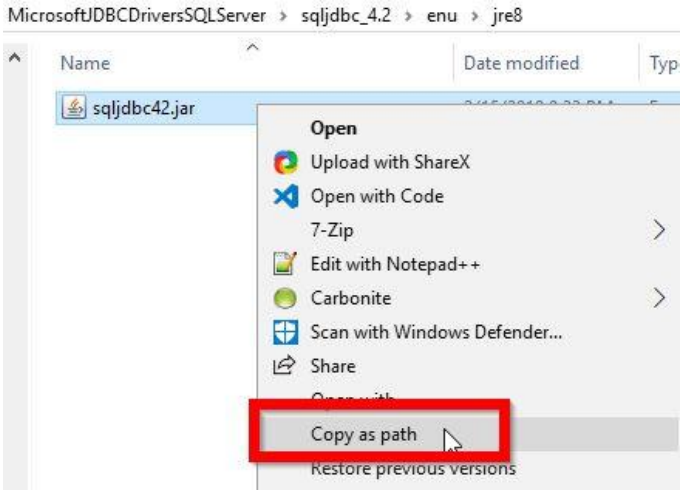
Una vez que queramos exportar los datos de SQL, hacemos uso de elasticsearch y logstash. Lo que se requiere es tener configurados las variables de entrono.

Vamos a menú del equipo y clicamos la edición de las variables de entorno.



Nos dirigimos a la variable “path”

Vamos a donde hayamos instalado el jdbc, le damos clic derecho y copiar como path. Y le pegamos ahí, en donde vayamos a configurar las variables de entorno,



Ahora, instalamos el driver, lo hacemos mediante la línea de comandos

```
.;"C:\misc\Java\Microsoft JDBC Drivers SQL Server\sqljdbc_4.2\enu\jre8\sqljdbc42.jar"
```

Una vez las tengamos tenemos que hacer una configuración:

Nos dirigimos al Logstash.conf y le añadimos en la parte del input el jdbc_concetion_string, el cual tendrá la conexión que se hace entre la base de datos y elasticsearch

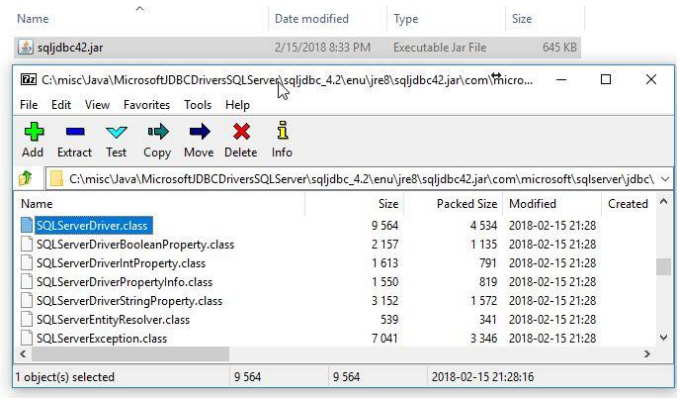
En la parte del output, hacemos referencia al localhost

```
input {
  jdbc {
    jdbc_connection_string => "jdbc:sqlserver://cc:1433;databaseName=StackExchangeCS;integratedSecurity=true;"
    jdbc_driver_class => "com.microsoft.sqlserver.jdbc.SQLServerDriver"
    jdbc_user => "xxx"

    statement => "SELECT * FROM Users"
  }
}

output {
  elasticsearch {
    hosts => ["localhost:9200"]
    index => "cs_users"
  }
}
```

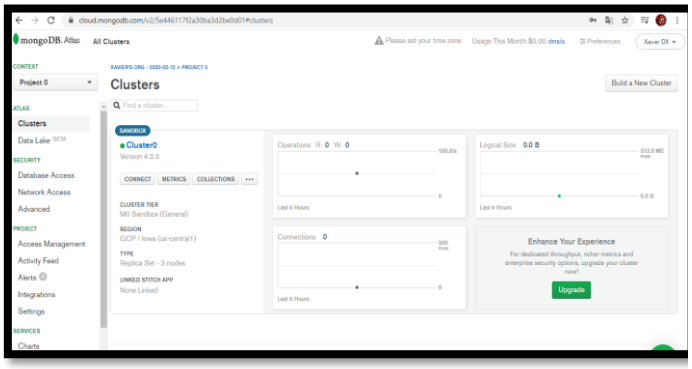
Luego de esto, tenemos que configurar la clase, la cual tendrá un componente que ayudará en la conexión



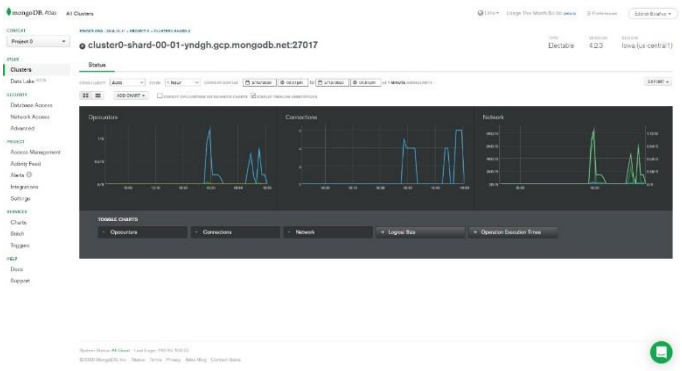
Entonces lo que hemos hecho hasta ahora es la conexión, pero para importar los datos… ¡Crear un clauter!

En este punto tuvimos un problema al utilizar kibana, lamentablemente el uso de hermmaientas como kibana involucra también tener corriendo el elasticseash. Eso, incluyendo algunos programas hacen que los recursos entren en conflicto.

En este punto decidimos hacer un replanteo de la convergencia de datos, y nos dimos cuenta que Mongo ofrece algunas herramientas que nos ayudan a hacer el análisis que necesitamos para este proyecto



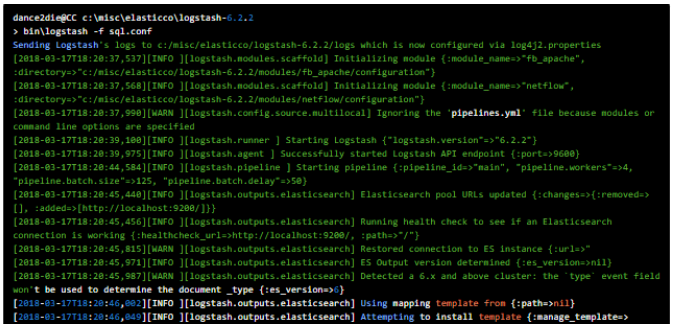
El Clúster inicialmente tendrá ese aspecto, pero ya cuando trabajemos con datos, se verá así:



Si ustedes, hubieran utilizado kibana tendría que editar un archivo .conf

```
bin\logstash -f sql.conf
```

Lo cual hará que la exportada sea exitosa:



El webscrapin consiste en la extracción de datos, de directamente de una página web, analizando su código html, con el uso de varias herramientas, como Selenium, librerías como py scraping, entre otros.

En nuestro caso, decidimos hacer todo con mongodb, por lo que los archivos SQL, los dejamos en standby y nos concentramos en las bases NOSQL, haciendo una cosecha de los datos, por cada temática planteada. Entonces, nos dirigimos a mongodb y ahí usamos lo que seria mongo Atlas.

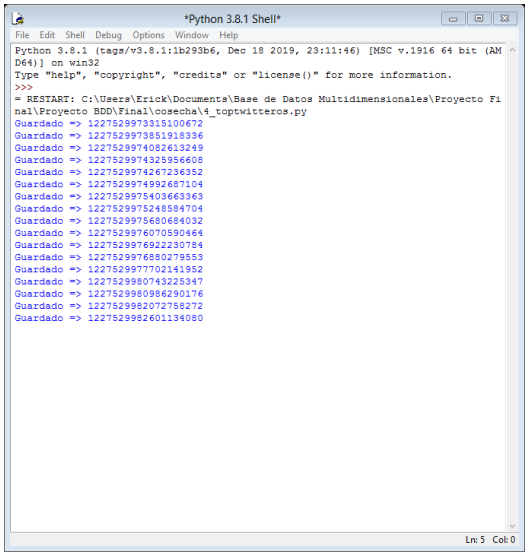
Mongo Atlas es la base de datos como servicio que permite implementar, utilizar y escalar una base de datos de MongoDB.
Ese servicio cuenta con una serie de herramientas que nos permiten hacer analítica de datos.
Una de esas herramientas es el mongo charts, que, al igual que suele hacer google chars, nos permite hacer visualizaciones de todos los datos que recopilamos

Usualmente, estos programas simulan la navegación de un humano en la World Wide Web ya sea utilizando el protocolo HTTP manualmente, o incrustando un navegador en una aplicación.

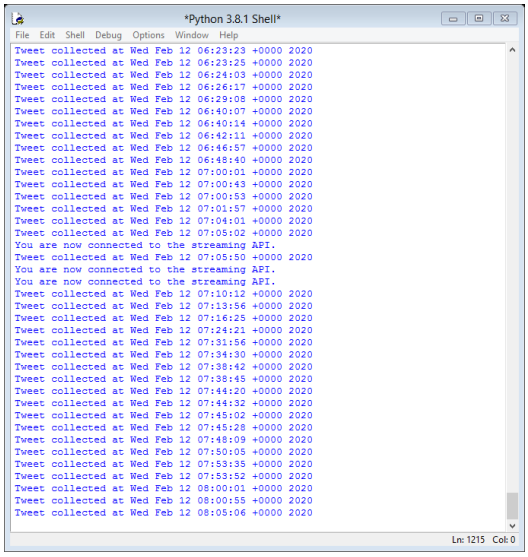
Ahora, vamos a trabajar con las bases de datos NoSQL

Utilizando los scripts de cosecha recopilábamos datos, y mandábamos directamente al clúster de MongoDB.

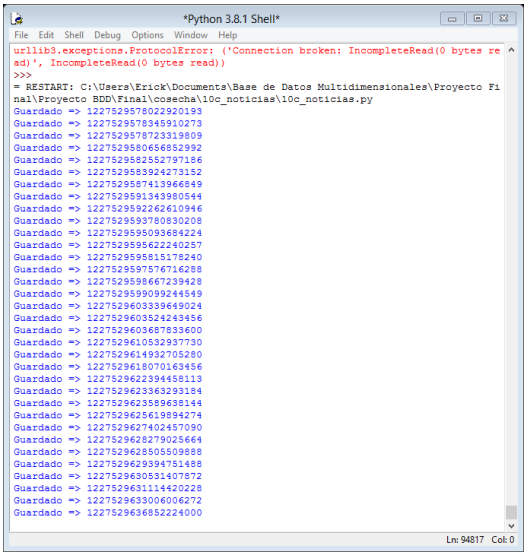
Apartado 4: Descarga desde un script en Python a MongoDB.



Apartado 9: Descarga desde un script en Python a MongoDB.



Apartado 10: Descarga desde un script en Python a MongoDB.



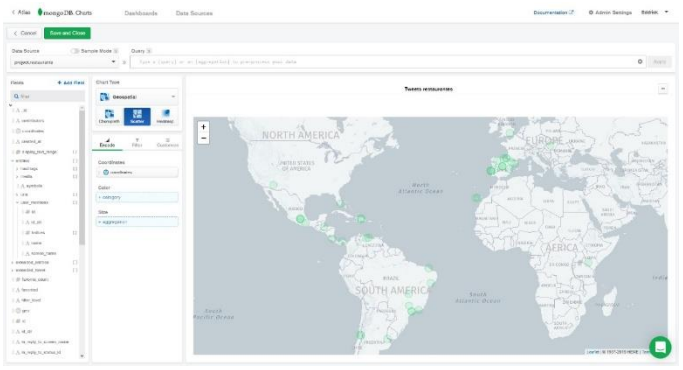
X. ANÁLISIS DE LA INFORMACIÓN

Para realizar un optimo análisis se realiza un mapeo en Elasticsearch en su apartado de “rest” que está en la banda superior

XI. VISUALIZACIÓN DE LA INFORMACIÓN

Para la visualización, Mongo ofrece varias opciones, entre ella la geolocalización.

La que verán a continuación es una geolocalización de “Eventos y noticias mundiales”. Y si vemos a detalle, podemos observar que las más producentes de estas noticias vienen de España, México, Venezuela, entre otras regiones de América



XII. RESULTADOS OBTENIDOS

XIII. CONCLUSIONES

MongoDB permite hacer análisis de datos fácilmente, es lo recomendable si el objetivo del curso es aprender a hacer análisis de datos, mediante la minería y otras técnicas de análisis.

Sin embargo, si el objetivo es aprender a usar herramientas, se recomienda que estas sean implementadas en computadores que tengan la capacidad.

Existe información de diferentes bases de datos, así como herramientas que nos permiten hacer un proceso de ETL para esos datos,

Los datos, una vez que son agrupados generan información, esa información nos permite hacer proyecciones. Ese es en principio el objetivo del proyecto. Y en todo caso, se cumple si tomamos en cuenta el objetivo general: la creación de la data wharehouse.

XIV. RECOMENDACIONES

Preguntar más acerca de la realización del proyecto con respecto a nuevas herramientas que puedan usarse y hacer mas sencilla la recopilación de datos.

Averiguar todo acerca de Google Charts para poder implementar más fácilmente.

XV. DESAFÍOS Y PROBLEMAS ENCONTRADOS

Instalar Kibana mostrando la interfaz, error de las migraciones de Elasticsearch a kibana.

Encontrar las bases de datos relacionales para cada caso de estudio correspondiente.

Seguir el cronograma trazado pese al resto de proyectos por hacer.

El daño de una laptop específicamente el daño del disco duro durante todo este proceso, y los datos recogidos de algunas temáticas se destinaron a perder lamentablemente.

Y el uso de webscraping no fue eficiente, nos faltó conocimiento de algunas herramientas de Python, así como el uso de sus librerías pero, sobre todo, tiempo.

XVI. ENLACE DE GITHUB DEL PROYECTO

https://github.com/EddRick96/FinalProject_BDDM_BHP.git.