



# ESCUELA POLITÉCNICA NACIONAL

## ESCUELA DE FORMACIÓN DE TECNÓLOGOS



### Integrantes:

- Miguel Cuenca
- Luis Jácome
- Kevin Morales
- Roberth Pincha
- Carlos Quel
- Sebastian Valencia
- Michael Valenzuela

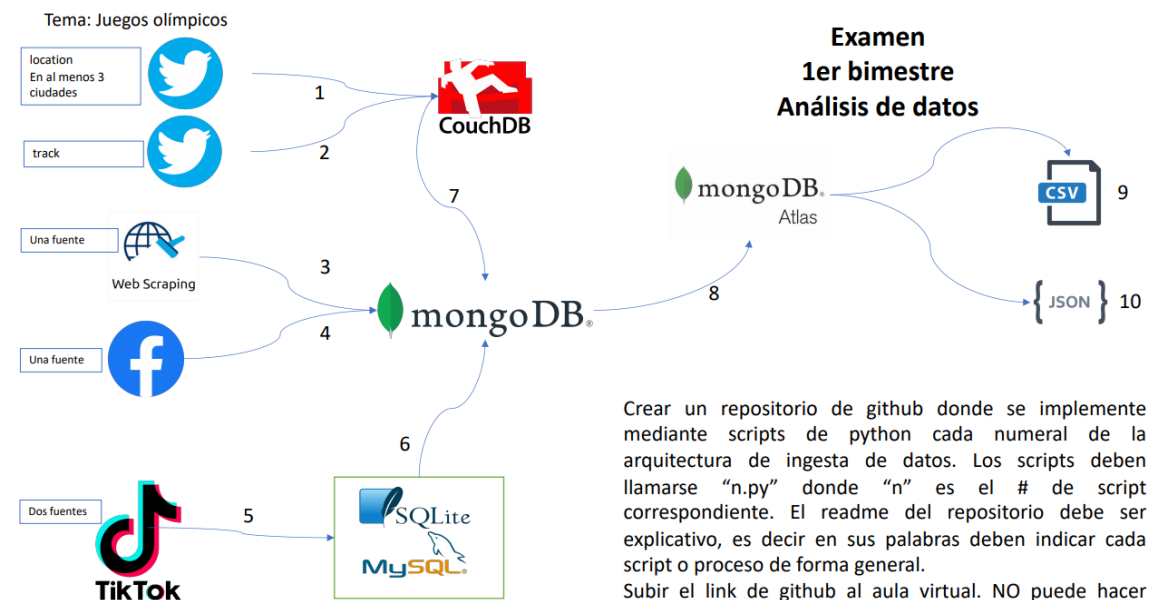
**Materia:** Análisis de Datos

**Grupo:** GR1

**Tema:** Deber Rapiminer

### Propósito:

- Implementar el flujo de datos propuesto en el examen, con RapidMiner.
- Puede obviar la conexión con la nube de MongoDB y realizarlo localmente.
- Explicar el proceso con capturas en un documento.
- Pueden trabajar en grupos, todos deben subir el documento.



## 1. Api Twitter

### Datos de Twitter a MongoDB Compass

Pasos:

Ejecutamos el script de Python que lo único que hace es realizar la extracción de datos de Twitter a través de una geolocalización de una primera ciudad del Ecuador por medio de las

respectivas credenciales que nos ofrece Twitter para este tipo de extracción de datos, y guarda en un archivo con extensión .csv.

```
SAVED\pymongo.results.InsertOneResult object at 0x00000200D1FAAD00=><{"created_at": "Wed Aug 25 23:15:04 +0000 2021", "id": "1430670088428982275", "id_str": "1430670088428982275", "text": "\u00a1Quiero comer corvichees con salrita y mayonesa!\n\u00d3ca href=\n\"http://twitter.com/download/iphone\" rel=\n\"nofollow\" \u003eTwitter for iPhone\u003c\n\"/>\n\"in_reply_to_status_id\":null,\"in_reply_to_status_id_str\":null,\"in_reply_to_user_id\":null,\"in_reply_to_user_id_str\":null,\"in_reply_to_screen_name\":null,\"user\":{\n\"id\":138137338,\"id_str\":\"138137338\",\"name\":\"Mel Ma\n\"screen_name\":\"mlismas12\",\"location\":\"Manta, Guayaquil, Ecuador\",\"url\":\"https://www.facebook.com/mlismas120/\", \"description\": \"Chola Pata Salada en el Guayas\", \"translator_type\": \"none\", \"protected\": false, \"verified\": false, \"followers_c\nount\": 298, \"friends_count\": 424, \"listed_count\": 0, \"favourites_count\": 3204, \"statuses_count\": 3310, \"created_at\": \"Fri Apr 26 05\n:25:38 +0000 2013\", \"utc_offset\": null, \"time_zone\": null, \"geo_enabled\": true, \"lang\": null, \"contributors_enabled\": false, \"is tr\nanslator\": false, \"profile_background_color\": \"000000\", \"profile_background_image_url\": \"http://abs.twimg.com/images/them\nes/theme1/bg.png\", \"profile_background_image_url_https\": \"https://abs.twimg.com/images/themes/theme1/bg.png\", \"prof\nile_background_tile\": false, \"profile_link_color\": \"E81C4F\", \"profile_sidebar_border_color\": \"000000\", \"profile_sidebar_fill_c\nolor\": \"000000\", \"profile_text_color\": \"000000\", \"profile_use_background_image\": false, \"profile_image_url\": \"http://pbs.twin\ns.com/profile_images/1365532342115332896/qcZLgr9_normal.jpg\", \"profile_image_url_https\": \"https://pbs.twimg.com/pro\nfile_images/1365532342115332896/qcZLgr9_normal.jpg\", \"profile_banner_url\": \"https://pbs.twimg.com/profile_banners/14\n38137338/1566322641\", \"default_profile\": false, \"default_profile_image\": false, \"following\": null, \"follow_request_sent\": null\n, \"notifications\": null, \"withheld_in_countries\": [], \"geo\": null, \"coordinates\": null, \"place\": {\n\"id\": \"01ee3b752f31d5cd\", \"url\": \"https://api.twitter.com/1.1/geo/id/01ee3b752f31d5cd.json\", \"place_type\": \"city\", \"name\": \"Manta, E\ncuador\", \"country_code\": \"EC\", \"country\": \"Ecuador\", \"bounding_box\": {\n\"type\": \"Polygon\", \"coordinates\": [[[-81.083528, -1.291199],\n[-81.083528, -0.927757], [-80.657451, -0.927757], [-80.657451, -1.291199]]], \"attributes\": {\n\"is_quote_s\natus\": false, \"quote_count\": 0, \"reply_count\": 0, \"retweet_count\": 0, \"favorite_count\": 0, \"entities\": {\n\"urls\": [], \"u\nser_mentions\": [], \"symbols\": [], \"favorited\": false, \"retweeted\": false, \"filter_level\": \"low\", \"lang\": \"es\", \"timestamp_ms\": 1629\n93304363\"}\n}
```

Ejecutamos otro script de Python que lo que hace es también realizar la extracción de datos de Twitter a través de una geolocalización de una segunda ciudad del Ecuador por medio de las respectivas credenciales que nos ofrece nuevamente Twitter para este tipo de extracción de datos, y guarda en un archivo con extensión .csv.

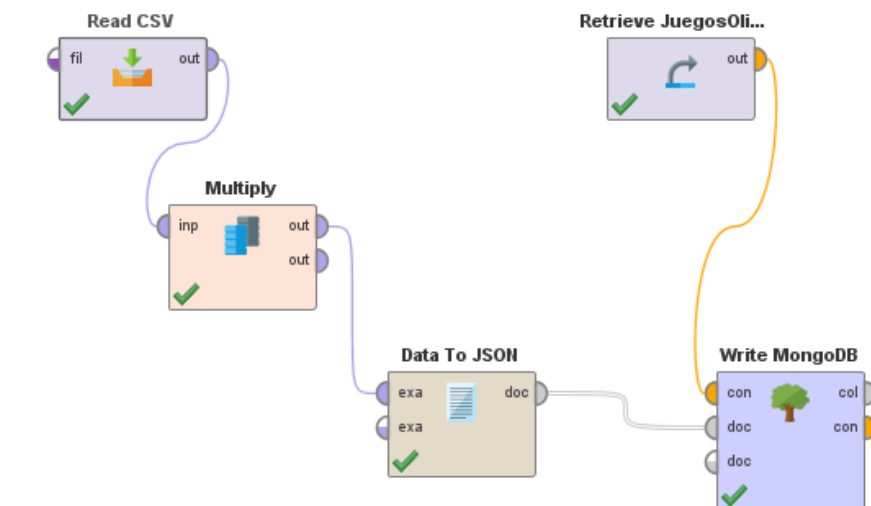
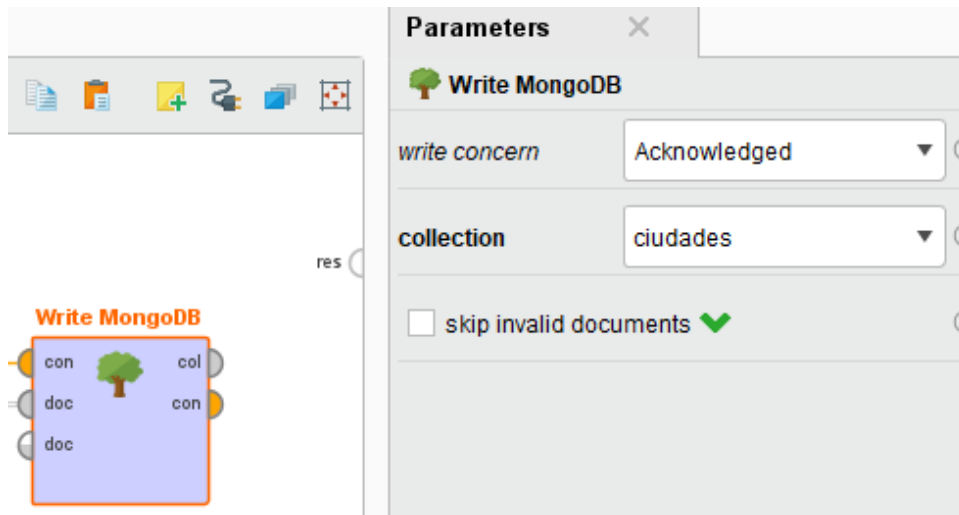
```
C:\Users\Andre\Desktop\PoliDeber3-2doBimestre\AnálisisDeDatos\Dataset\Twitter>python twitter1.2.py
SAVED\pymongo.results.InsertOneResult object at 0x00000189184EEF40=><{"created_at": "Wed Aug 25 23:03:50 +0000 2021", "id": "1430667262147665924", "id_str": "1430667262147665924", "text": "\"Uno de mis libros favoritos, 'Love Story', de Erich Segal, se\n\u00d3fiala que la protagonista adoraba a Ethebeates. Eso me\n\u00d226 https://t.co/m1t5Z7cn10\", \"source\": \"\n\"http://twitter.com/download/android\" rel=\n\"nofollow\" \u003eTwitter for Android\u003c\n\"/>\n\"in_reply_to_status_id\":null,\"in_reply_to_status_id_str\":null,\"in_reply_to_user_id\":null,\"in_reply_to_user_id_str\":null,\"in_reply_to_screen_name\":null,\"user\":{\n\"id\":158585905,\"id_str\":\"158585905\", \"name\": \"Jose\n\"screen_name\":\"Mor\n\"location\":\"Guayaquil, Ecuador\", \"url\": \"https://www.facebook.com/josexavier.mor\n\"description\": \"Dr. en Med. Veterinaria & Zootecnia. #Geminiano. Liberal. Algo existencialista, esteta, insomne, grunge, sn\nob, hereje y\n\"followers_count\": 2857, \"friends_count\": 1234, \"listed_count\": 34, \"favourites_count\": 140456, \"statuses_count\": 120918, \"created_at\": \"Wed Jun 23 02:45:46 +0000 2010\", \"utc_offset\": null, \"time_zone\": null, \"geo_enabled\": true, \"lang\": null, \"contributors_ena\nbled\": false, \"is translator\": false, \"profile_background_color\": \"FAEBF3\", \"profile_background_image_url\": \"http://abs.t\ns.twimg.com/images/themes/theme1/bg.gif\", \"profile_background_image_url_https\": \"https://abs.twimg.com/images/themes/th\neme1/bg.gif\", \"profile_background_tile\": false, \"profile_link_color\": \"8B8E7E\", \"profile_sidebar_border_color\": \"000000\", \"pr\nofile_sidebar_fill_color\": \"000000\", \"profile_text_color\": \"000000\", \"profile_image_url\": \"http://abs.twimg.com/pro\nfile_images/1406021428865126400/CEHjPkP_normal.jpg\", \"profile_image_url_https\": \"https://pbs.twimg.com/pro\nfile_images/1406021428865126400/CEHjPkP_normal.jpg\", \"profile_banner_url\": \"https://pbs.twimg.com/pro\nfile_banners/158585905/1601226451\", \"default_profile\": false, \"default_profile_image\": false, \"following\": null, \"follow\nrequest_sent\": null, \"notifications\": null, \"withheld_in_countries\": [], \"geo\": null, \"coordinates\": null, \"place\": {\n\"id\": \"009924a469d7ace1\", \"url\": \"https://api.twitter.com/1.1/geo/id/009924a469d7ace1.json\", \"place_type\": \"city\", \"name\": \"Guayaquil\n\", \"full_name\": \"Guayaquil, Ecuador\", \"country_code\": \"EC\", \"country\": \"Ecuador\", \"bounding_box\": {\n\"type\": \"Polygon\", \"coordinates\": [[[-80.467605, -3.063716], [-80.467605, -1.962274], [-79.714781, -1.962274], [-79.714781, -3.063716]]], \"attributes\": {\n\"cont\nributors\": null, \"quoted_status_id\": \"1430605591462952966\", \"quoted_status_id_str\": \"1430605591462952966\", \"quoted_status\": \"\n\"created_at\": \"Wed Aug 25 18:58:47 +0000 2021\", \"id\": \"1430605591462952966\", \"id_str\": \"1430605591462952966\", \"text\": \"\n\"}
```

Ejecutamos el otro script también de ciudades de Python que lo único que también hace es realizar la extracción de datos de Twitter a través de una geolocalización de una tercera ciudad del Ecuador por medio de las respectivas credenciales que nos ofrece nuevamente Twitter para este tipo de extracción de datos, y guarda en un archivo con extensión .csv.

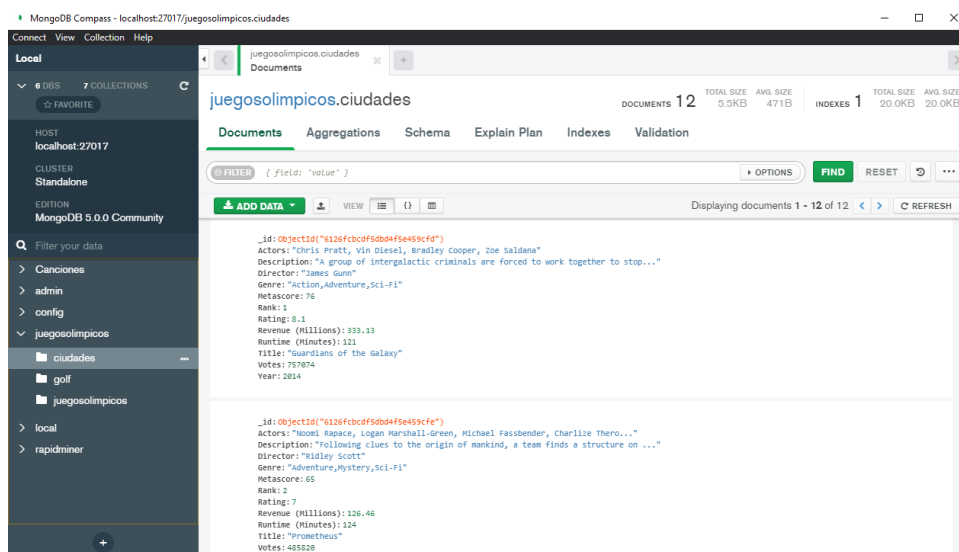
```
C:\Users\Andre\Desktop\PoliDeber3-2doBimestre\AnálisisDeDatos\Dataset\Twitter>python twitter1.1.py
SAVED\pymongo.results.InsertOneResult object at 0x0000020020027C00=><{"created_at": "Wed Aug 25 23:07:54 +0000 2021", "id": "1430668287768207360", "id_str": "1430668287768207360", "text": "\"La Alcaid\n\"display_text_range\": [0,140], \"source\": \"\n\"http://twitter.com/download/iphone\" rel=\n\"nofollow\" \u003eTwitter for iPhone\u003c\n\"/>\n\"in_reply_to_status_id\":null,\"in_reply_to_status_id_str\":null,\"in_reply_to_user_id\":null,\"in_reply_to_user_id_str\":null,\"in_reply_to_screen_name\":null,\"user\":{\n\"id\":1631290380,\"id_str\":\"1631290380\", \"name\": \"\n\"screen_name\":\"Ag\n\"location\":\"Manta - Ecuador\", \"url\": \"http://epan.gob.ec\", \"des\ncription\": \"Cuenta oficial de la empresa p\n\"followers_count\": 5760, \"friends_count\": 123, \"listed_count\": 24, \"favourites_count\": 1043, \"statuses_count\": 7174, \"created_at\": \"Mon Jul 29 21:40:15 +0000 2013\", \"utc_offset\": null, \"time_zone\": null, \"geo_enabled\": true, \"lang\": null, \"contributors_enabled\": false, \"is translator\": false, \"profile_background_color\": \"C0DEED\", \"profile_background_image_url\": \"http://abs.t\ns.twimg.com/images/themes/theme1/bg.png\", \"profile_background_image_url_https\": \"https://abs.twimg.com/images/themes/th\neme1/bg.png\", \"profile_background_tile\": false, \"profile_link_color\": \"0084B4\", \"profile_sidebar_border_color\": \"FFFFFF\", \"pr\nofile_sidebar_fill_color\": \"DDEEFF\", \"profile_text_color\": \"333333\", \"profile_use_background_image\": false, \"profile_image_u\nrl\": \"http://pbs.twimg.com/profile_images/1410716869550538753/55yc37_n_normal.jpg\", \"profile_image_url_https\": \"https://pbs.twimg.com/profile_images/1410716869550538753/55yc37_n_normal.jpg\", \"profile_banner_url\": \"https://pbs.twimg.com/profile_banners/1631290380/1628638590\", \"default_profile\": false, \"default_profile_image\": false, \"following\": null, \"follow_request_sent\": null, \"notifications\": null, \"withheld_in_countries\": [], \"geo\": null, \"coordinates\": null, \"place\": {\n\"id\": \"01ee3b752f31d5cd\", \"url\": \"https://api.twitter.com/1.1/geo/id/01ee3b752f31d5cd.json\", \"place_type\": \"city\", \"name\": \"Manta\", \"full_name\": \"Manta, Ecuador\", \"country_code\": \"EC\", \"country\": \"Ecuador\", \"bounding_box\": {\n\"type\": \"Polygon\", \"coordinates\": [[[-81.083528, -1.291199], [-81.083528, -0.927757], [-80.657451, -0.927757], [-80.657451, -1.291199]]], \"attributes\": {\n\"contribu\nors\": null, \"is_quote_status\": false, \"extended_tweet\": {\n\"full_text\": \"La Alcaid\n\"mu\ny pronto, a los\n\"frente al exhot\n\"salau text_range\": [0,225], \"entities\": {\n\"text\": \"\n\"}
```

Ejecutamos nuevo script de Python que lo único que hace es realizar la extracción de datos de Twitter a través de palabras claves buscadas directamente de Twitter (track) por medio de las respectivas credenciales que nos ofrece nuevamente Twitter para este tipo de extracción de datos, y guarda en un archivo con extensión .csv.





Al final podemos comprobar en nuestra BD los datos que han sido trasladados con éxito.



## 2. API Web Scarping

### Datos de Web Scraping(Python) a MongoDB Compass

Pasos:

Ejecutamos el script de Python, su funcionamiento consta de realizar una minería de datos dentro de un sitio web, para conseguir ciertos datos se deben especificar ciertos criterios en la estructura que identifican las etiquetas contenedoras de esos datos y posteriormente su traslado a una lista.

```
In [21]: response = requests.get('https://olympics.com/tokyo-2020/es/deportes/')
response2 = requests.get('https://resultados.as.com/resultados/juegos_olimpicos/medallero/')
soup = BeautifulSoup(response.content, "lxml")
soup2 = BeautifulSoup(response2.content, "lxml")

Sport=[]
Country=[]

#se escape la etiqueta span
post_sports = soup.find_all("h2", class_="tk-disciplines_title")
post_countries = soup2.find_all("span", class_="team-inline")

#for para limpiar
for element in post_sports:
    #print(element)
    element=str(element)
    limpio=str(element[find_1st(element, '>')+1:find_2nd(element, '<')])
    #print (limpio)
    Sport.append(limpio.strip())

#for para limpiar
for element in post_countries:
    #print(element)
    element=str(element)
    limpio=str(element[find_1st(element, '>')+1:find_2nd(element, '<')])
    #print (limpio)
    Country.append(limpio.strip())

print(Sport)
print("=====")
print(Country)

['Atletismo', 'Badminton', 'Baloncesto', 'Baloncesto 3x3', 'Balonmano', 'Béisbol', 'Boxeo', 'Ciclismo BMX freestyle', 'Ciclismo o BMX racing', 'Ciclismo de montaña', 'Ciclismo en pista', 'Ciclismo en ruta', 'Escalada deportiva', 'Esgrima', 'Fútbol', 'Gimnasia artística', 'Gimnasia rítmica', 'Gimnasia: trampolín', 'Golf', 'Halterofilia', 'Hípica', 'Hockey', 'Judo', 'Karate', 'Lucha', 'Natación', 'Natación artística', 'Natación en aguas abiertas', 'Pentatlón moderno', 'Piragüismo en eslalon', 'Piragüismo en sprint', 'Remo', 'Rugby', 'Saltos', 'Skateboarding', 'Surf', 'Taekwondo', 'Tenis', 'Tenis de mesa', 'Tiro', 'Tiro con arco', 'Triatlón', 'Vela', 'Voleibol', 'Voleibol playa', 'Waterpolo']
=====
['Estados Unidos', 'China', 'Japón', 'Reino Unido', 'Comité Olímpico Ruso', 'Australia', 'Países Bajos', 'Francia', 'Alemania', 'Italia', 'Canadá', 'Brasil', 'Nueva Zelanda', 'Cuba', 'Hungria', 'Corea del Sur', 'Polonia', 'República Checa', 'Kenia', 'Noruega', 'Jamaica', 'España', 'Suecia', 'Suiza', 'Dinamarca', 'Croacia', 'Irán', 'Serbia', 'Bélgica', 'Bulgaria', 'Eslovenia', 'Uzbekistán', 'Georgia', 'China Taipei', 'Turquía', 'Grecia', 'Uganda', 'Ecuador', 'Israel', 'Irlanda', 'Qatar', 'Kosovo', 'Bahamas', 'Ucrania', 'Bielorrusia', 'Rumania', 'Venezuela', 'India', 'Hong Kong', 'Filipinas', 'Eslovaquia', 'Sudáfrica', 'Austria', 'Egipto', 'Indonesia', 'Portugal', 'Etiopía', 'Túnez', 'Estonia', 'Tailandia', 'Fidji', 'Letonia', 'Bermudas', 'Puerto Rico', 'Marruecos', 'Colombia', 'Azerbaiyán', 'República Dominicana', 'Armenia', 'Kirguizistán', 'Mongolia', 'Argentina', 'San Marino', 'Jordania', 'Malasia', 'Nigeria', 'Turkmenistán', 'Macedonia', 'Namibia', 'Lituania', 'Bahrein', 'Arabia Saudita', 'Kazajistán', 'México', 'Finlandia', 'Kuwait', 'Costa de Marfil', 'Ghana', 'Siria', 'Burkina', 'Granada', 'Moldavia', 'Botswana', 'Rusia', 'Corea del Norte', 'Vietnam', 'Participante Olímpico Independiente', 'Tajikistán', 'Singapur', 'Argelia', 'Burundi', 'Nigeria', 'Trinidad y Tobago', 'Emiratos Árabes']
```

Posteriormente a la creación de las listas, debemos convertir estos datos en la estructura de dataset para ser identificados por los gestores de base de datos.

```
In [26]: #Trasladar La Lista a un dataframe
bd1=pd.DataFrame(Sport)
#Trasladar La Lista a un dataframe
bd2=pd.DataFrame(Country)
```



```
In [28]: bd2
```

```
Out[28]:
```

	0
0	Estados Unidos
1	China
2	Japón
3	Reino Unido
4	Comité Olímpico Ruso
...	...
99	Argelia
100	Burundi
101	Niger
102	Trinidad y Tobago
103	Emiratos Árabes

104 rows x 1 columns

## Uso de Mysql Workbench

Para trasladar los dataset en mongodb a través de Rapidminer, nosotros hemos decidido elegir el gestor de BD Mysql para realizar la vinculación. Dentro del script es necesario tener en nuestro FrameWork las librerías para tener los comandos necesarios para la conexión.

```
In [29]: pip install PyMySQL
```

```
Collecting PyMySQL
  Downloading PyMySQL-1.0.2-py3-none-any.whl (43 kB)
Installing collected packages: PyMySQL
Successfully installed PyMySQL-1.0.2
Note: you may need to restart the kernel to use updated packages.

WARNING: You are using pip version 21.1.2; however, version 21.2.4 is available.
You should consider upgrading via the 'c:\users\pc-vict\appdata\local\programs\python\python39\python.exe -m pip install --upgr
ade pip' command.
```

```
In [33]: import pymysql
```

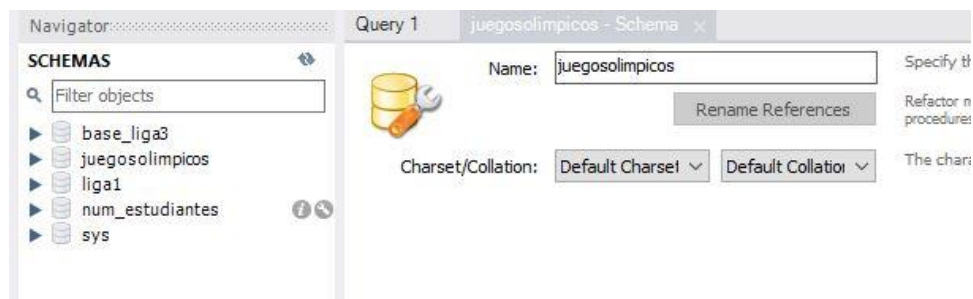
```
In [34]: pip install sqlalchemy
```

```
Collecting sqlalchemy
  Note: you may need to restart the kernel to use updated packages.
  Downloading SQLAlchemy-1.4.23-cp39-cp39-win_amd64.whl (1.5 MB)
Requirement already satisfied: greenlet!=0.4.17 in c:\users\pc-vict\appdata\local\programs\python\python39\lib\site-packages (f
rom sqlalchemy) (1.1.0)
Installing collected packages: sqlalchemy
Successfully installed sqlalchemy-1.4.23

WARNING: You are using pip version 21.1.2; however, version 21.2.4 is available.
You should consider upgrading via the 'c:\users\pc-vict\appdata\local\programs\python\python39\python.exe -m pip install --upgr
ade pip' command.
```

```
In [35]: from sqlalchemy import create_engine
```

En Mysql debemos crear la base de datos de la cual recibirá dichos contenidos.



De regreso en nuestro script debemos integrar un script para realizar esta conexión de Python a Mysql.

```
In [37]:
sqlEngine = create_engine('mysql+pymysql://root:T27op870mn012@127.0.0.1/juegosolimpicos', pool_recycle=3600)
dbConnection = sqlEngine.connect()

try:
    frame = bd1.to_sql("Deportes", dbConnection, if_exists='fail');
except ValueError as vx:
    print(vx)
except Exception as ex:
    print(ex)
else:
    print("Table %s created successfully");
finally:
    dbConnection.close()

c:\users\pc-vict\appdata\local\programs\python\python39\lib\site-packages\pandas\io\sql.py:1444: UserWarning: The provided table name 'Deportes' is not found exactly as such in the database after writing the table, possibly due to case sensitivity issues. Consider using lower case table names.
warnings.warn(msg, UserWarning)
```

```
In [39]:
sqlEngine = create_engine('mysql+pymysql://root:T27op870mn012@127.0.0.1/juegosolimpicos', pool_recycle=3600)
dbConnection = sqlEngine.connect()

try:
    frame = bd2.to_sql("Paises", dbConnection, if_exists='fail');
except ValueError as vx:
    print(vx)
except Exception as ex:
    print(ex)
else:
    print("Table %s created successfully");
finally:
    dbConnection.close()

Table %s created successfully

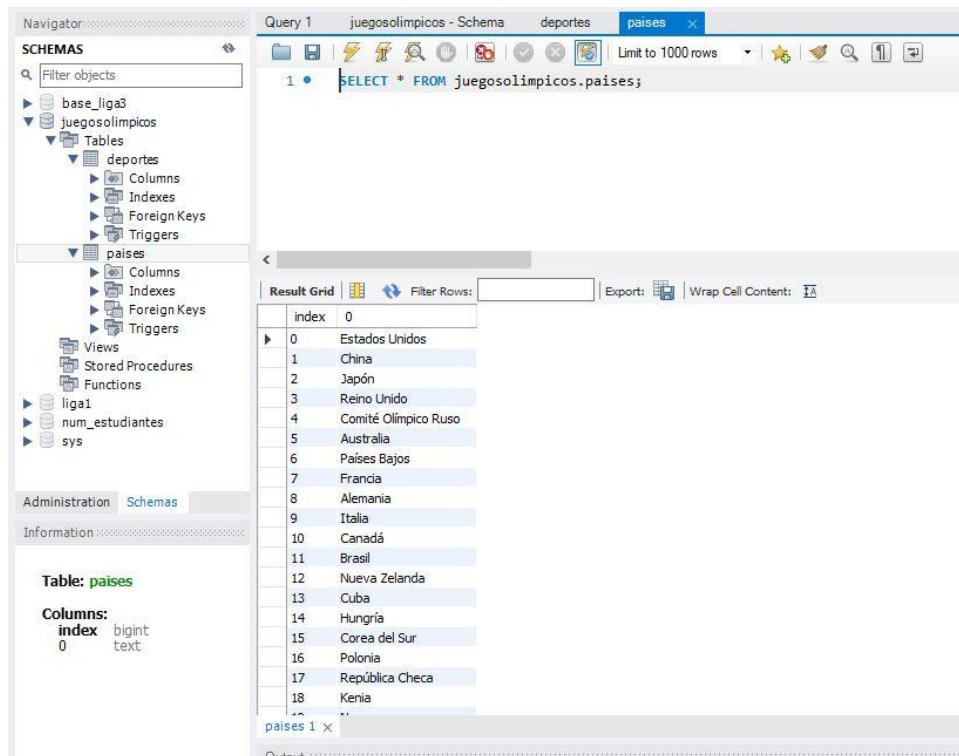
c:\users\pc-vict\appdata\local\programs\python\python39\lib\site-packages\pandas\io\sql.py:1444: UserWarning: The provided table name 'Paises' is not found exactly as such in the database after writing the table, possibly due to case sensitivity issues. Consider using lower case table names.
warnings.warn(msg, UserWarning)
```

Una vez terminado cada proceso nos dirigimos a nuestra BD para verificar si los dataset han sido integrados.

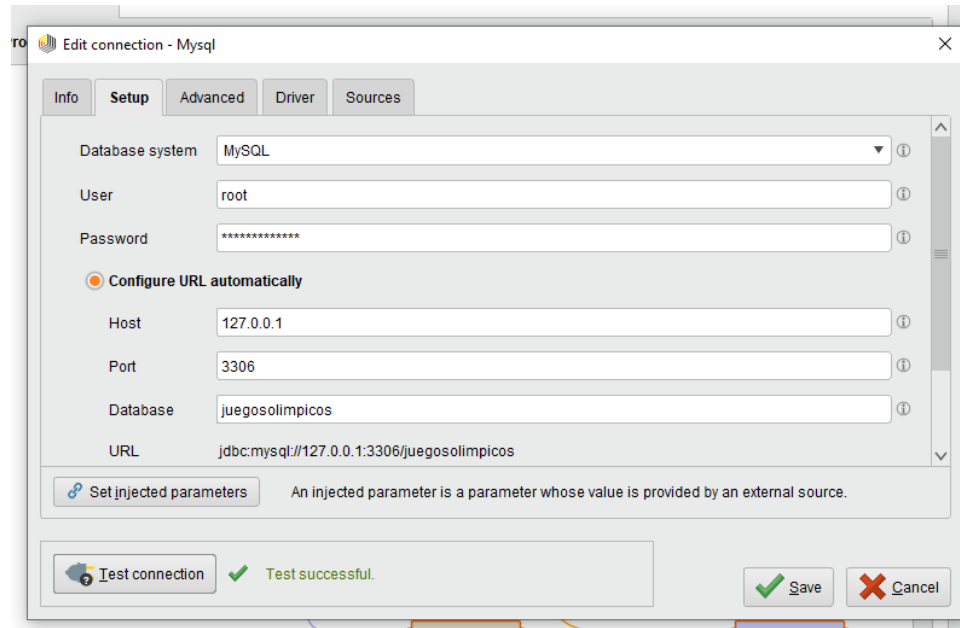
The screenshot shows a database management interface with the following components:

- Navigator:** Displays the database schema hierarchy. The 'juegosolimpicos' database is expanded, showing tables like 'base\_liga3', 'deportes', 'liga1', 'num\_estudiantes', and 'sys'.
- Query 1:** The SQL query is `SELECT * FROM juegosolimpicos.deportes;`
- Result Grid:** Shows the data for the 'deportes' table. The first column is 'index' (0 to 18) and the second column is 'text' (various sports like Atletismo, Bádminton, Baloncesto, etc.).
- Table: deportes:** A summary of the table structure:
 

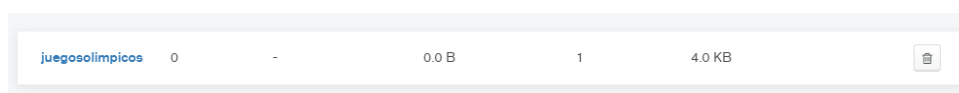
Columns:	index	text
0	bigint	



En rapiminer, creamos una nueva conexión que apunte a la base de datos de nuestra preferencia de la cual será importada a rapidminer para realizar todo tipo de evaluación.

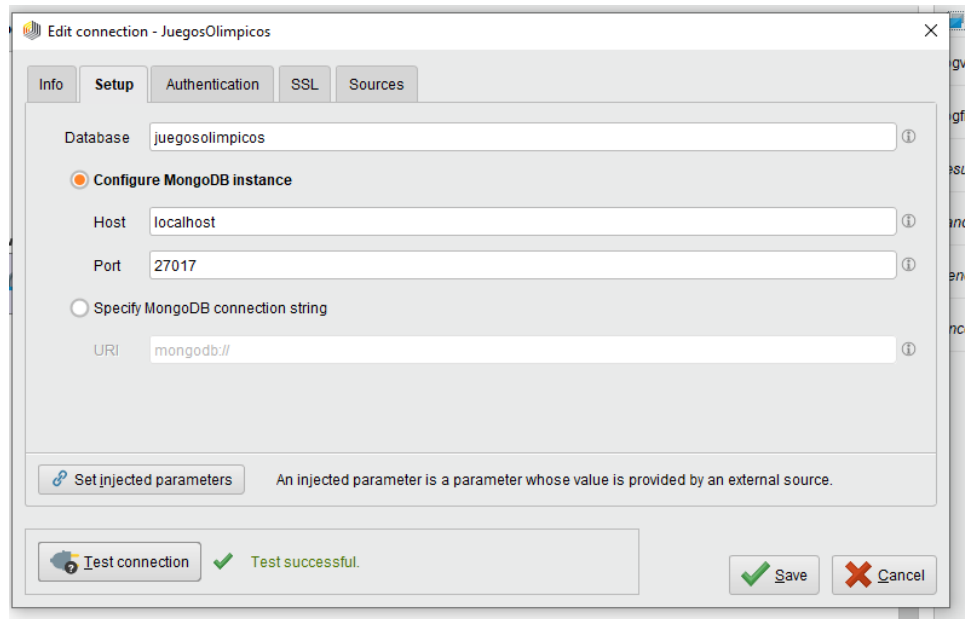


Además, en mongo compass creamos la colección que servirá como repositorio para nuestras datasets.

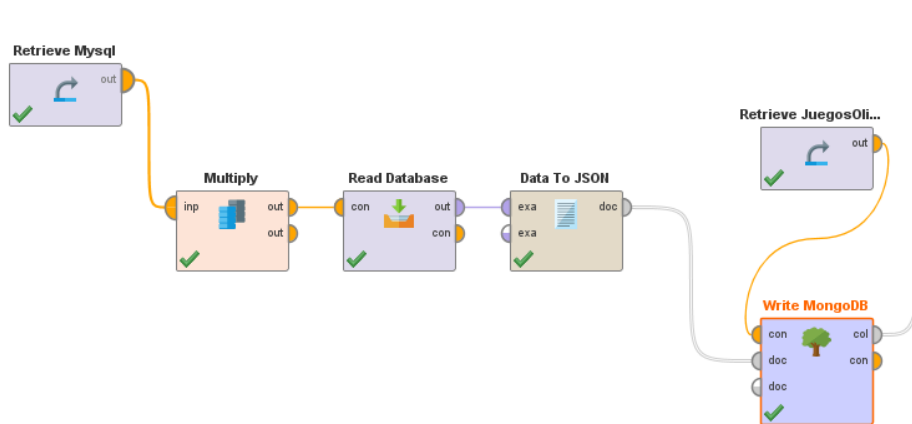




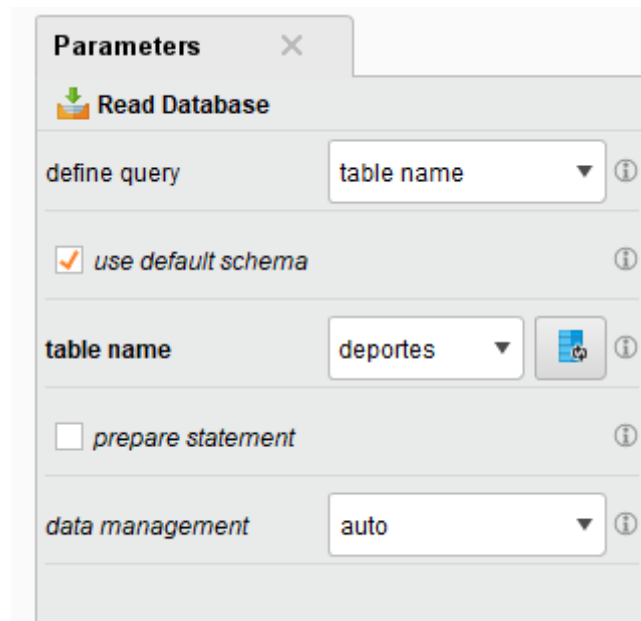
En Rapidminer, creamos una nueva conexión con MongoDB de la cual nos servirá para instanciarla y poder enviar los datos en la BD de nuestra preferencia.



Una vez establecida las configuraciones de nuestras BD lo sigue es realizar el modelo con los componentes de rapidminer para pasar los dataset.

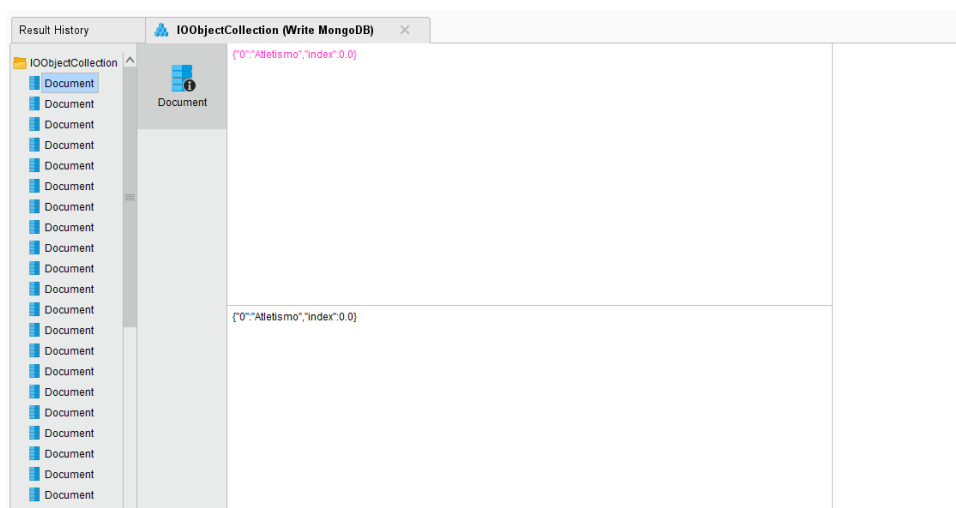


Con el componente Read Database podremos seleccionar una de las tablas que han sido establecidas en MYSQL.

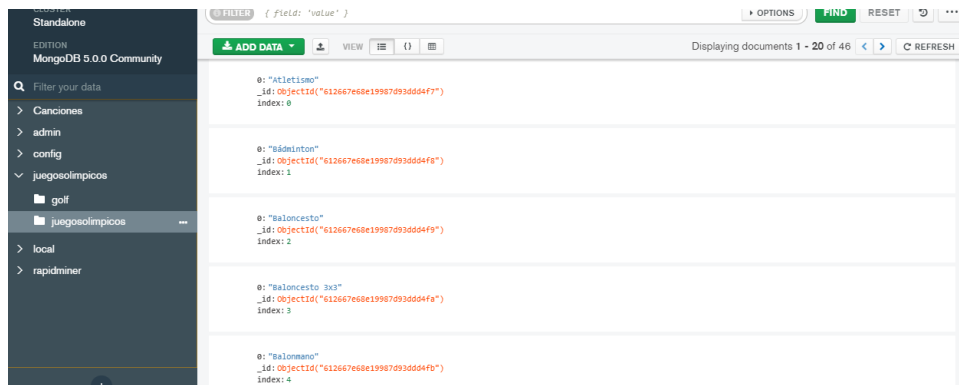


Luego el componente Data to Json, convertira estos datos en una estructura Json para que mongo pueda identificarla y poder registrar en nuestra colección.

Si todo ha salido bien podemos observar como se genran la estructuras Json con los datos de Mysql.



Y posteriormente a ello podemos verificar en MongoDB si todo ha salido bien.



### 3. API Facebook

## Datos de Facebook a MongoDB Compass

Pasos:

Ejecutamos el script de Python que lo único que hace es realizar la extracción de datos de Facebook a través de palabras claves como en este caso ha sido “nintendo”, y guarda en un archivo con extensión .csv.

[illegible]

Creamos la base de datos “rapidminer” para que pueda ser leída y reconocida en el programa RapidMiner para su respectiva conexión.

rapidminer	4.0KB	1	1
------------	-------	---	---

Localizamos al archive con extensión .csv y lo importamos al programa RapidMiner lo cual no presenta problemas o errores. Una vez leído continuamos con su finalización de columnas, entre otros.

### Specify your data format

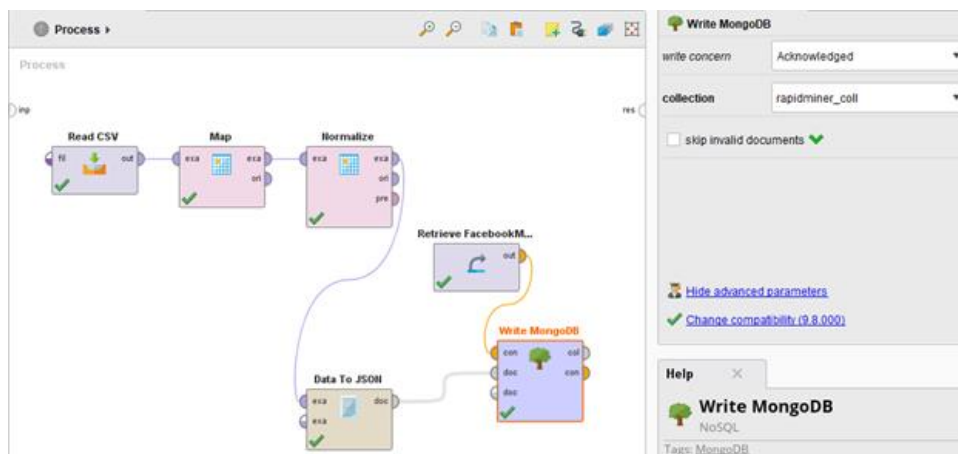
☒ Header Row   
 Start Row   
 Column Separator

File Encoding   
 Escape Character   
 Decimal Character

☒ Use Quotes   
☐ Trim Lines  
☒ Skip Comments

1	_id	comments	date	id	likes	post_url	reactions	shares	texto
2	6125ba37...	28	16298262...	44329298...	418	https://fac...	null	20	My Ninten...
3	6125ba3f...	1356	16297308...	44294111...	5118	https://fac...	null	2436	On this da...
4	6125ba50...	23	16295657...	44236454...	238	https://fac...	null	0	Creators f...
5	6125ba58...	196	16291512...	44087340...	826	https://fac...	null	48	Itâ€™s an...
6	6125ba5f...	200	16286004...	43889578...	1084	https://fac...	null	217	A new Indi...
7	6125ba69...	37	16282877...	43778509...	252	https://fac...	null	0	Kit & Kryst...
8	6125ba7e...	28	16282735...	43773800...	207	https://fac...	null	14	Lunchabl...
9	6125ba8a...	677	16282669...	43771216...	3078	https://fac...	null	936	SUPER N...
10	6125ba9f...	62	16281128...	43717276...	331	https://fac...	null	29	July 2021 ...
11	6125bab6...	437	16279349...	43656243...	1544	https://fac...	null	316	Super Mar...

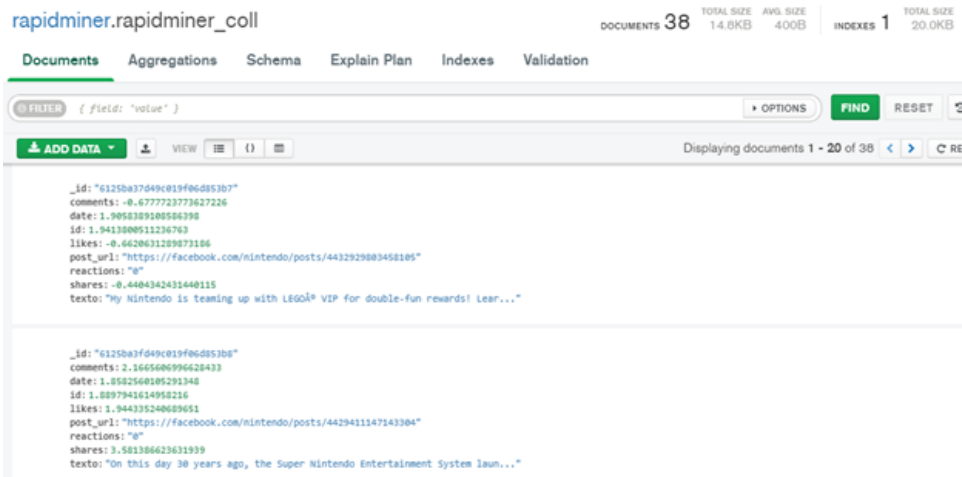
Arrastramos todos los operadores necesarios para la respectiva conexión con MongoDB Compass, incluso realizando un map y una normalización para la representación de los datos. También, ya creada la conexión con MongoDB en la base de datos “rapidminer”, continuamos a unir sin ningún problema. Finalmente, especificamos en el operador *Write MongoDB* la respectiva colección en la cual guardará todos los datos.



Una vez ejecutado y comprobado que no existen problemas o errores visualizamos la base de datos de MongoDB Compass en su colección adjuntada lo cual podemos observar que todos los datos han sido guardados de manera exitosa.

Collection Name ^	Documents	Avg. Document Size	Total Document Size	Num. Indexes	Total Index Size
rapidminer_coll	38	400.0 B	14.8 KB	1	4.0 KB

Una visualización más perfecta se la puede realizar con los respectivos datos guardados, mejores vistos en formato json que fue parseado directamente en el programa RapidMiner.



#### 4. Api Tiktok

- **Parte 5**

## Extracción de datos desde TikTok

### Fuente 1 Arigameplays

1. Después de instalar TikTok Scraper, procedo a ejecutar el comando:

```
tiktok-scraper user arigameplays -n 100 -t csv --session sid_tt=asd13123123123ad
```

Para obtener un archivo csv de la tiktok AriGameplays.

## Fuente 2 Chikybombomreal

- Después de instalar TikTok Scraper, procedo a ejecutar el comando:

```
tiktok-scraper user chikybombomreal -n 100 -t csv --session sid_tt=asdasd13123123123adasda
```

Para obtener un archivo csv de la tiktok Chikybombomreal.

```
File Edit Selection View Go Run Terminal Help chibombomreal_1629909159351.csv - TikTok - Visual Studio Code

EXPLORER
  X chibombomreal_1629...
  OPEN EDITORS
    X chibombomreal_1629...
  TIKTOK
    node_modules
    trend
    argameplays_162990890541...
    autoplayruebs_16269982099...
    caritas126_1626998118940.csv
    chibombomreal_16299091...
    chin_1626998539575.csv
    el_uar_1_1626998438876.csv
    jallynegeida_162699811862...
    lajuilet87_1626998483352.csv
    marypooz0_1626998251526.c...
  package-lock.json
  package.json
  polluxus_1627691526932.csv
  risapinea_1626998525828.csv
  rllap_1626998281415.csv
  sterenoficial_1626998349794...
  tiktokData.db

chibombomreal_1629909159351.csv
1  "id","secretID","text","createTime","authorMeta.id","authorMeta.secid","author
2  "7000200110361545980","7000200110361545980","El avión me dejó en Puerto Rico
3  "6998671608709664805","6998671608709664805","Esto es una bendición wow 🙏 que e
4  "7000156319806278918","7000156319806278918","Circuito en motor en arepublicadom
5  "P.O.BOK 228901 MIAMI,FL,33222","https://p16-sign-va.tiktokcdn.com/tos-maliva-
6  "6998671608709664805","6998671608709664805","Esto es una bendición wow 🙏 que e
7  "P.O.BOK 228901 MIAMI,FL,33222","https://p16-sign-va.tiktokcdn.com/tos-maliva-
8  "6998620713859042565","6998620713859042565","FELIZZZZ EN MO PAÍS REPUBLICA DOMIN
9  "P.O.BOK 228901 MIAMI,FL,33222","https://p16-sign-va.tiktokcdn.com/tos-maliva-
10 "6997947412031622406","6997947412031622406","BUENASBUENAS HOY ME SIENTO COMO LA
11 "P.O.BOK 228901 MIAMI,FL,33222","https://p16-sign-va.tiktokcdn.com/tos-maliva-
12 "6997927091257855238","6997927091257855238","En unos minutos estaré conectada e
13 "P.O.BOK 228901 MIAMI,FL,33222","https://p16-sign-va.tiktokcdn.com/tos-maliva-
14 "6997244547986181382","6997244547986181382","MIL gracias aorlandoflorida por
15 "P.O.BOK 228901 MIAMI,FL,33222","https://p16-sign-va.tiktokcdn.com/tos-maliva-
16 "6997061482898787590","6997061482898787590","Gloria a Dios encontré un restauran
17 "P.O.BOK 228901 MIAMI,FL,33222","https://p16-sign-va.tiktokcdn.com/tos-maliva-
18 "6996709904216661765","6996709904216661765","BUENASBUENAS 🙏🙏🙏,1629084671,"

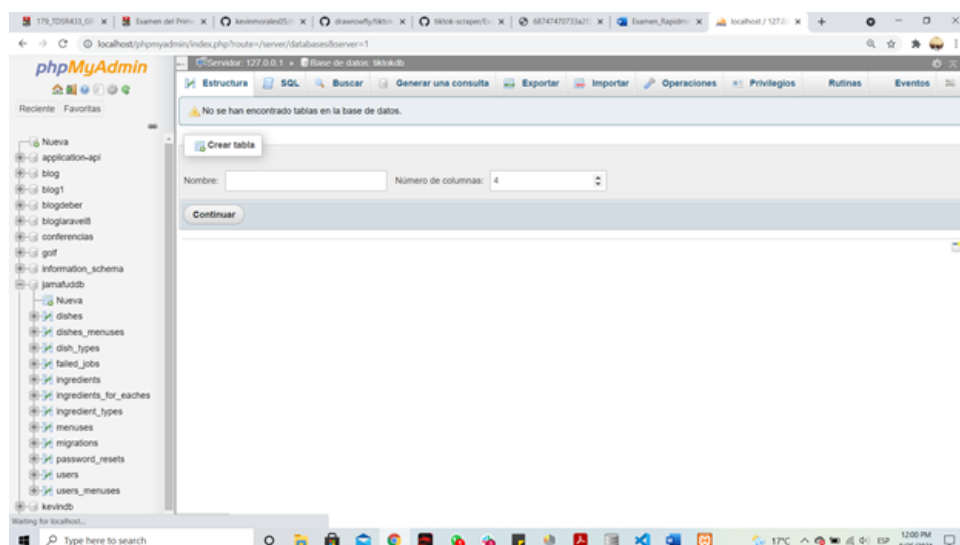
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL
kvede@KevINPSDeveloper: ~/Documents/TikTok
$ tiktok-scraper user argameplays -n 100 -t csv --session sid_tt=asdasd1312312312adasde
| TikTok Scraper Started
(process:18032): GLib-GIO-WARNING **: 11:28:24.851: Unexpectedly, UWP app '3138AweZlp.AweZip 1.4.8.0
x86_ffd30ambc3c' (APID '3138AweZlp.AweZip_ffd30ambc3c1App') supports 168 extensions but has no
verbs
CSV path: C:\Users\kvede\Documents\TikTok\argameplays_1629908905415.csv

kvede@KevINPSDeveloper: ~/Documents/TikTok
$ tiktok-scraper user chibombomreal -n 100 -t csv --session sid_tt=asdasd1312312312adasde
| TikTok Scraper Started
(process:27132): GLib-GIO-WARNING **: 11:32:38.816: Unexpectedly, UWP app '3138AweZlp.AweZip 1.4.8.0
x86_ffd30ambc3c' (APID '3138AweZlp.AweZip_ffd30ambc3c1App') supports 168 extensions but has no
verbs
CSV path: C:\Users\kvede\Documents\TikTok\chibombomreal_1629909159351.csv

kvede@KevINPSDeveloper: ~/Documents/TikTok
$
```

- **Parte 6**

Creo una base de datos en phpMyAdmin.

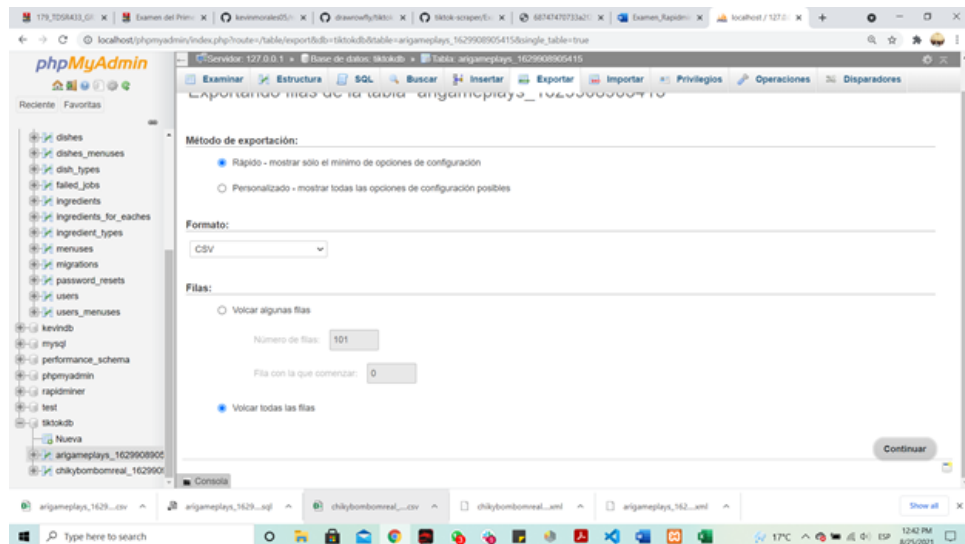


Después escojo el archivo CSV y lo importo.

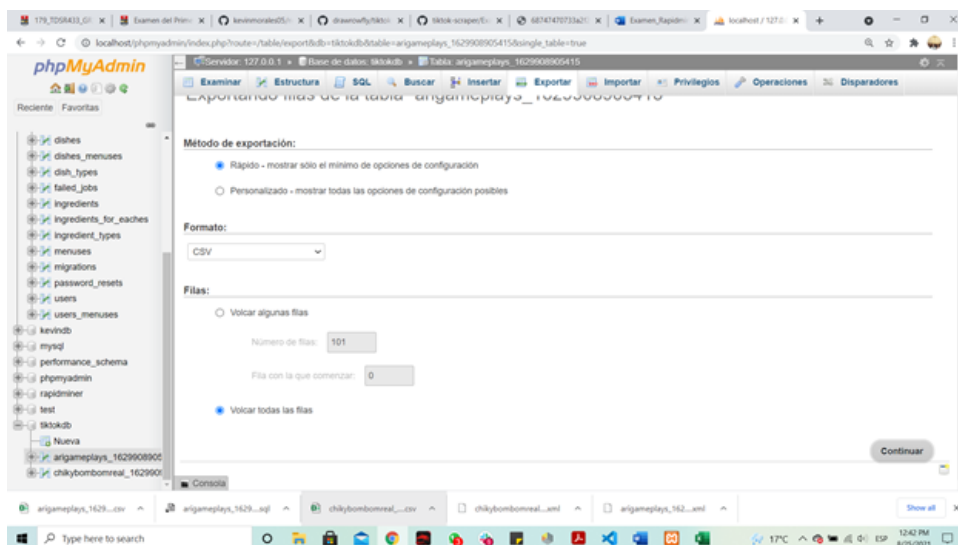




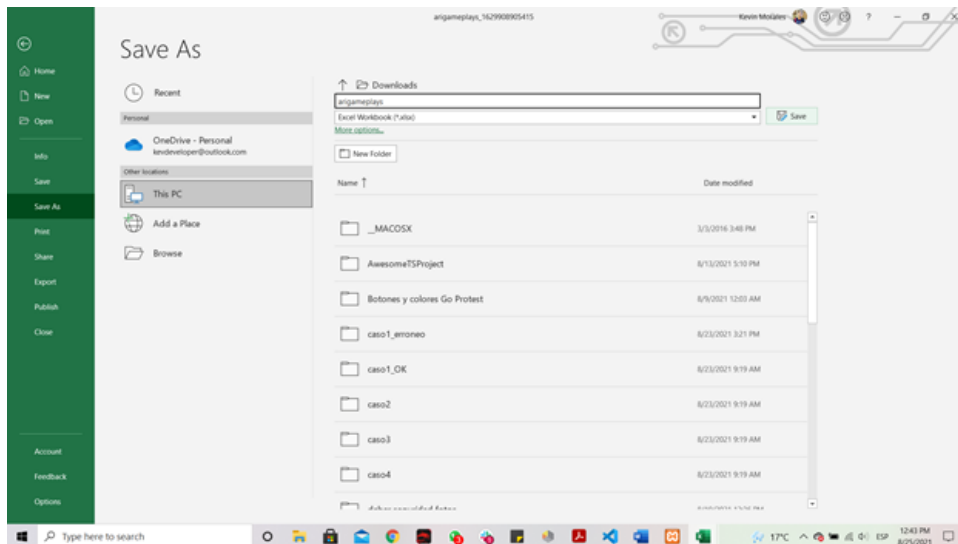
1. Exporto en formato CSV, exporto la tabla de Chikybombomreal de la misma manera, el objetivo de esto fue para realizar una limpieza previa de datos.



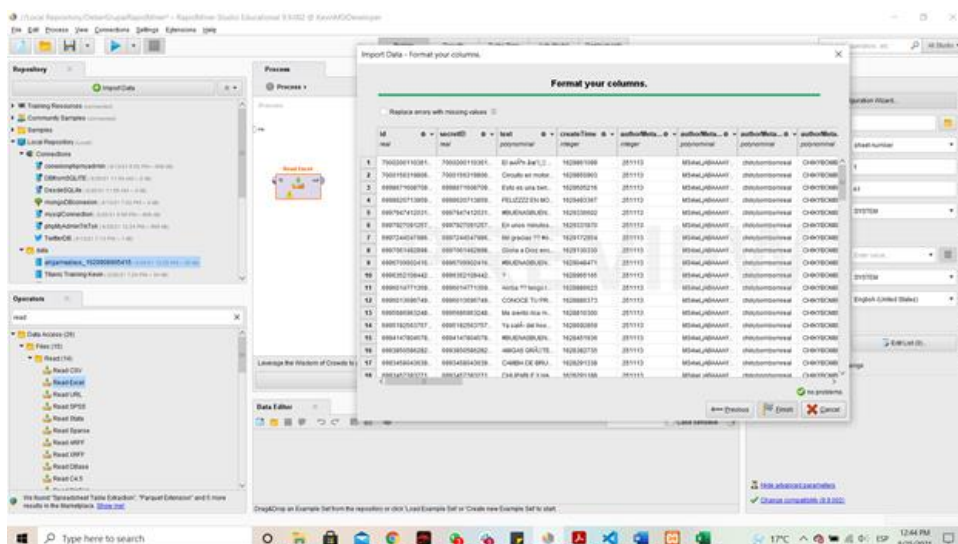
2. Luego de tener los archivo, se procede a abrir las tablas en archivos de Excel para limpiar la fila que esta llena de las columnas enumeradas, lo cual produce errores en RapidMiner.



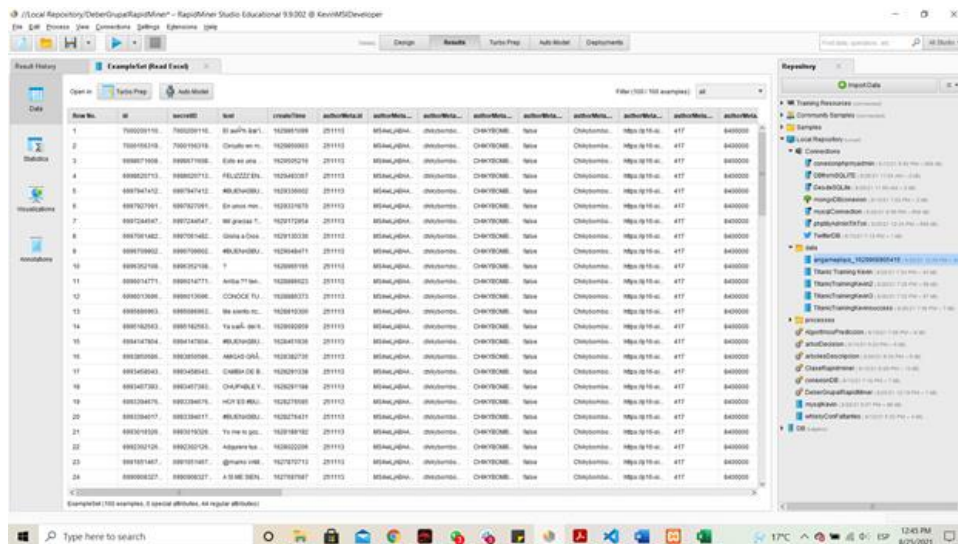
3. Después procedo a guardar como archivo Excel para luego usar un Operador READ EXCEL de RapidMiner.



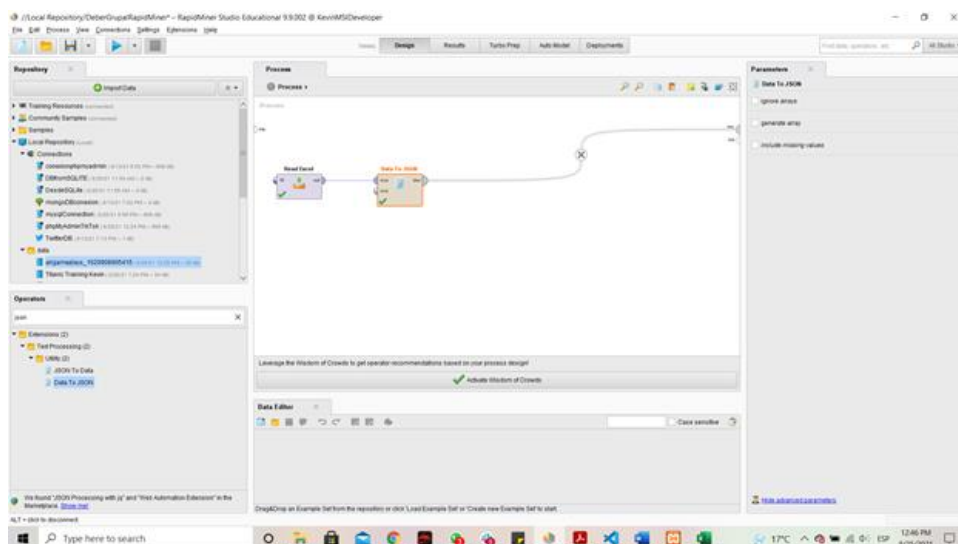
4. Con el operador Read Excel, obtengo toda la información de la tabla en RapidMiner. En este ejemplo se muestra la lectura de la tabla Chikybombomreal.



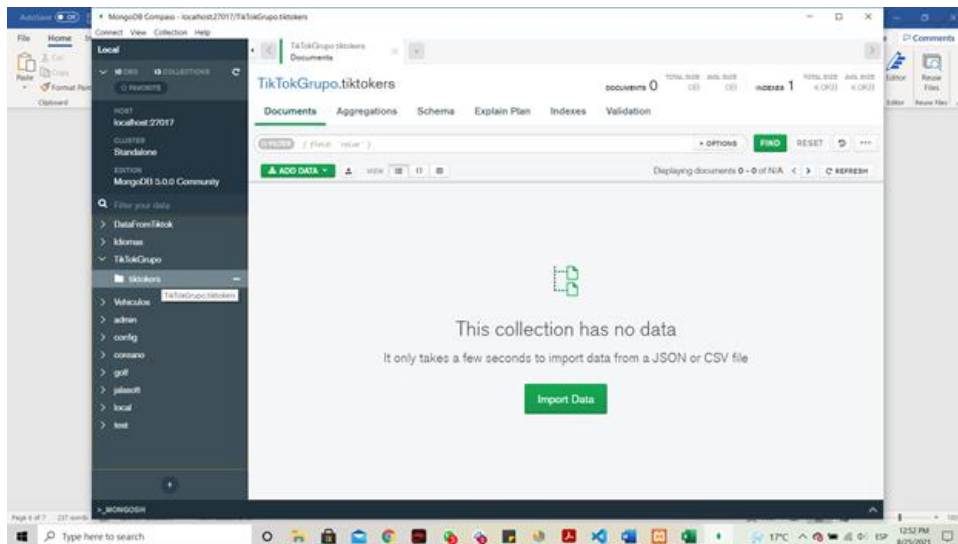
5. Luego al presionar PLAY se verifica la conexión en RapidMiner. Se realizo lo mismo con la tabla Arigameplays.



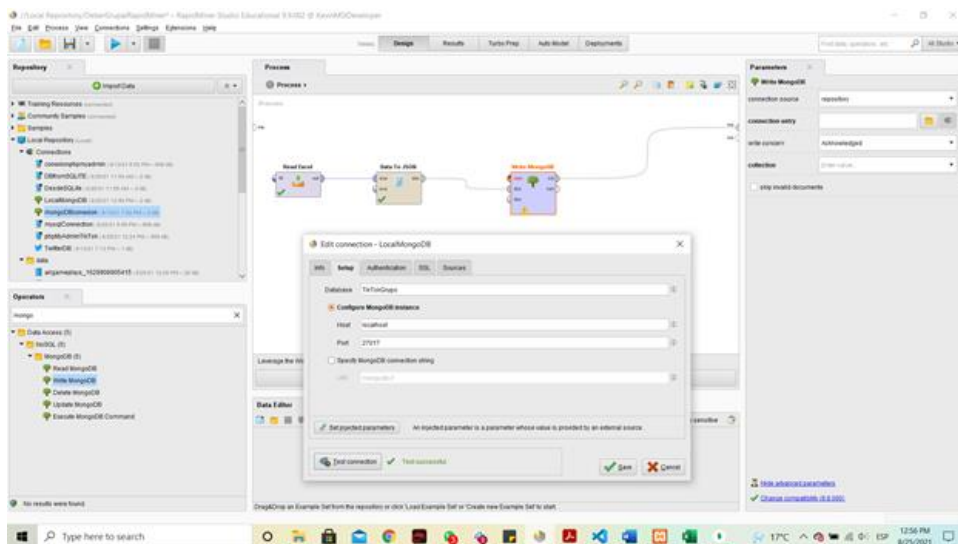
6. Convierto todos los datos a JSON, para luego mandarlos a MongoDB Compass.



7. Creo una base de datos en MongoDB en el servidor local para almacenar la información.

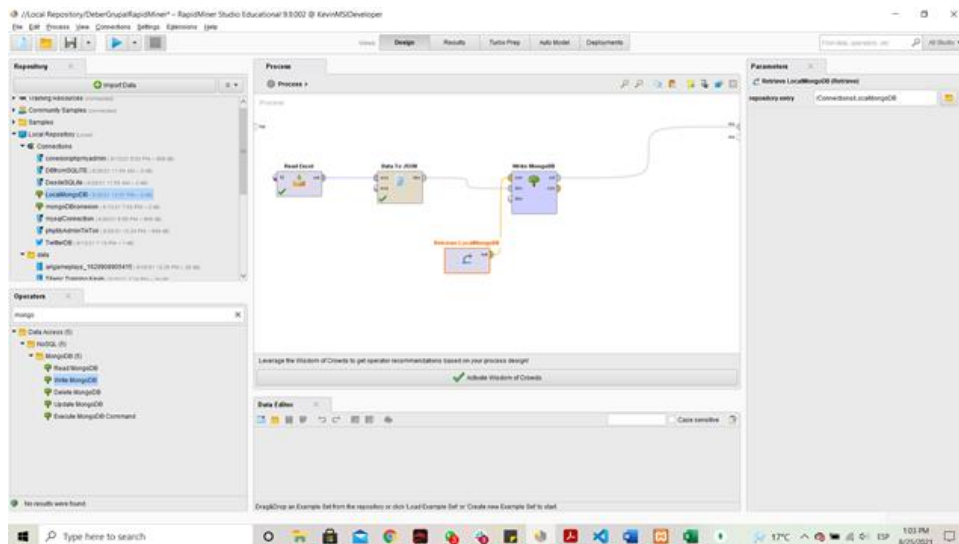


8. Creo una nueva conexión con MongoDB y coloco el nombre de la base de Datos creada.

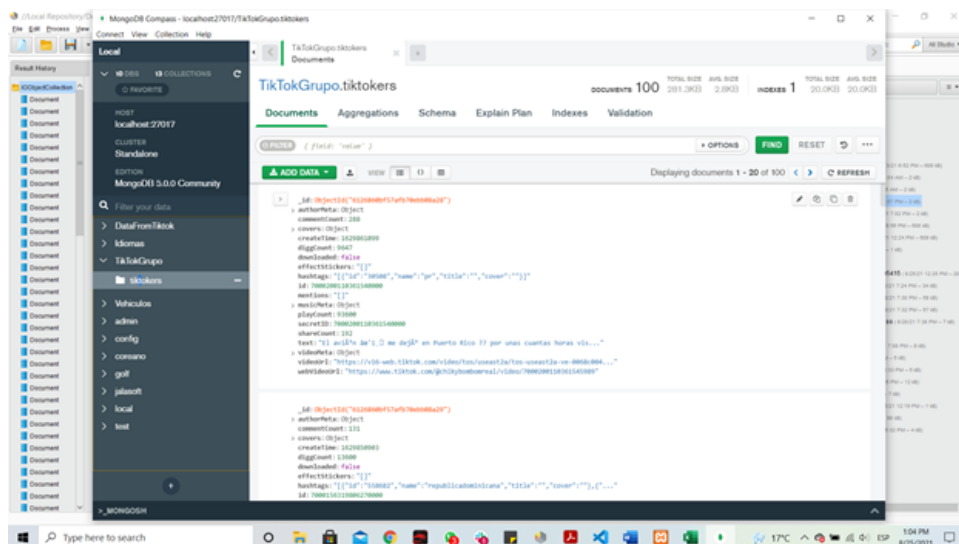


9. Conecto la conexión creada y conecto el operador Write To MongoDB, escojo la colección tiktokers.

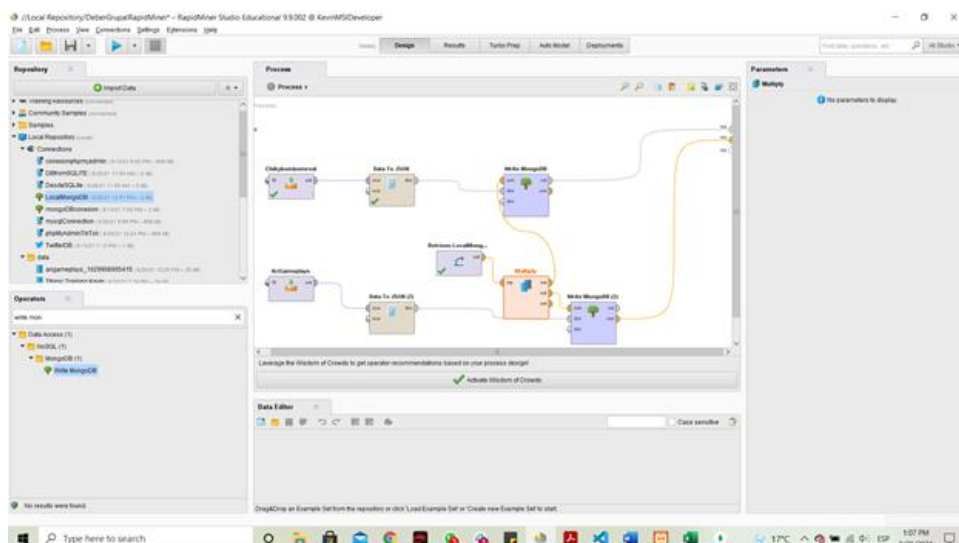




Con esta configuración se puede apreciar los datos en MongoDB Compass. Lo mismo se realizó con la otra tabla.

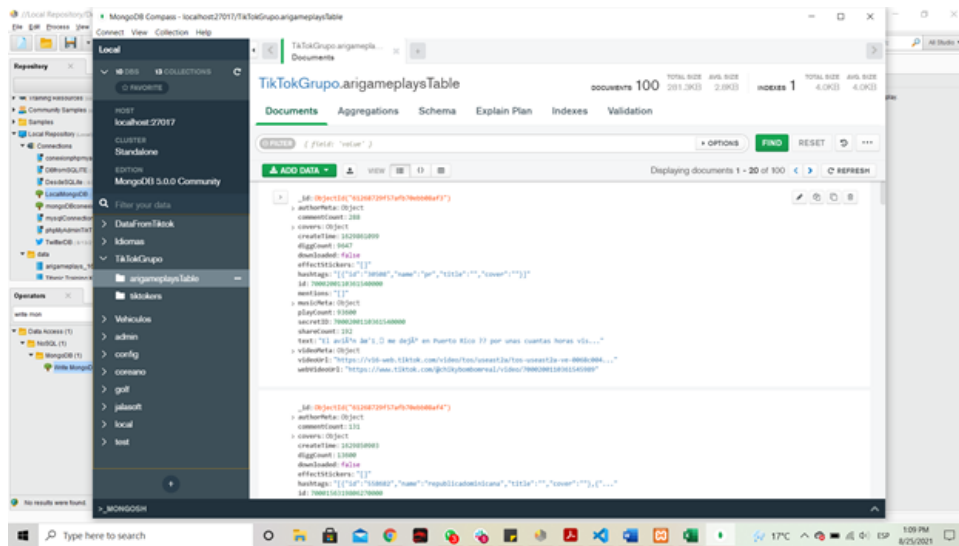


Para escribir la información de Arigameplays, creo una colección nueva en MongoDB Compass, luego procedo a realizar las siguientes conexiones de operadores:



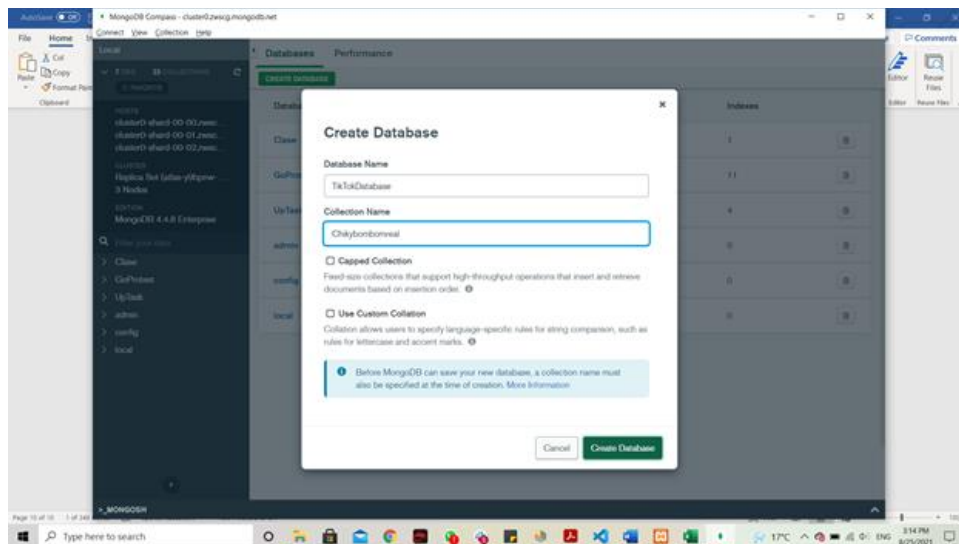


Obteniéndose estos resultados:



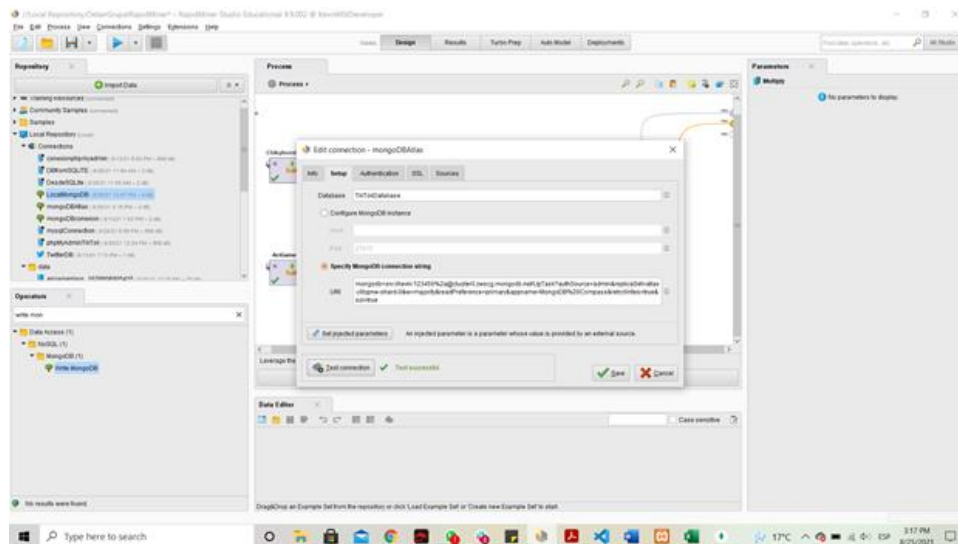
## • Parte 8

1. Para conectar a MongoDB Atlas, se procede a crear una base de datos en el cluster creado previamente:

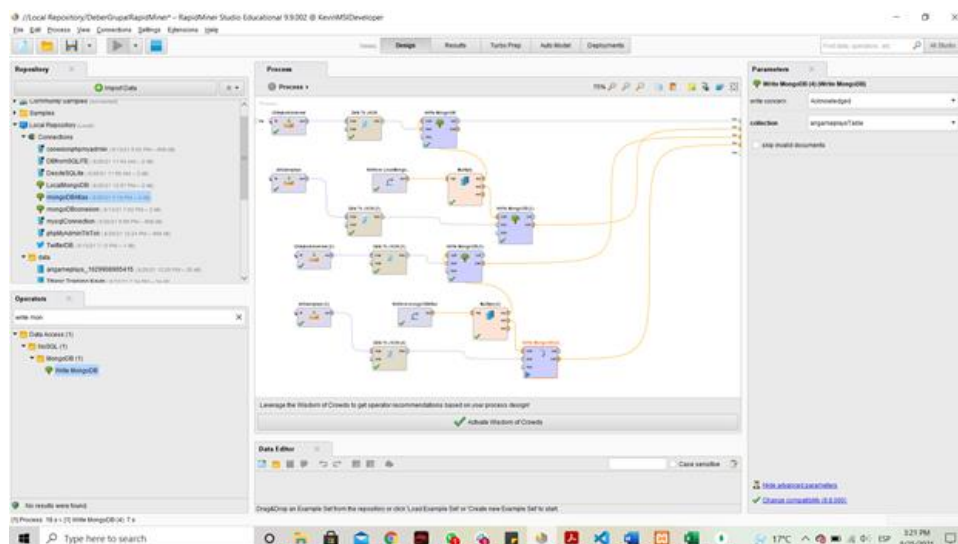


Se realiza el mismo procedimiento para la colección con los datos de AriGameplays.

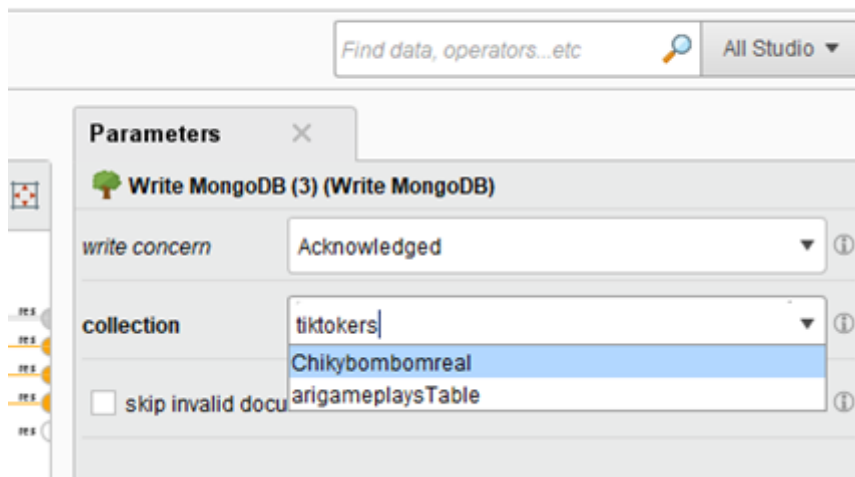
2. Creo una nueva conexión a MongoDB Atlas en RapidMiner.



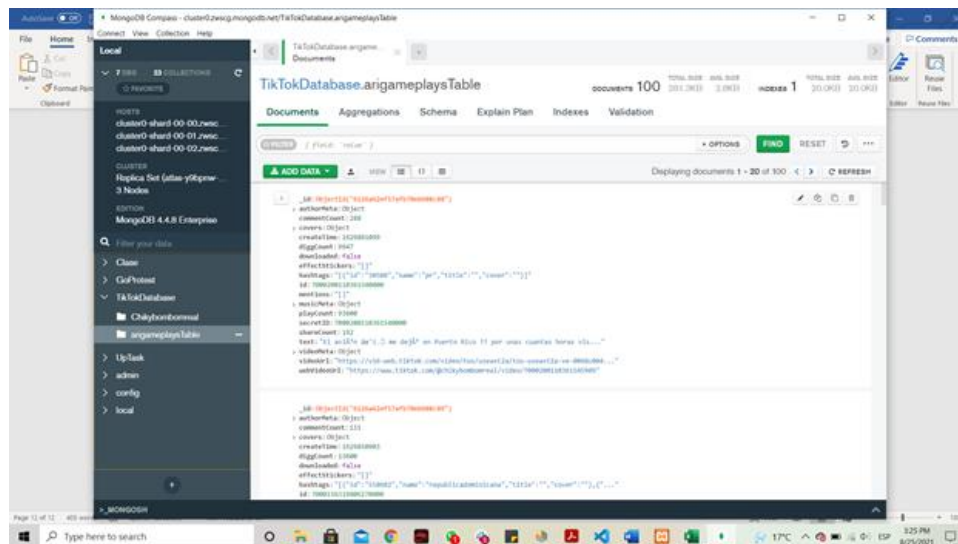
- Usando operadores, se procede a pasar las tablas al cluster en MongoDB Atlas.



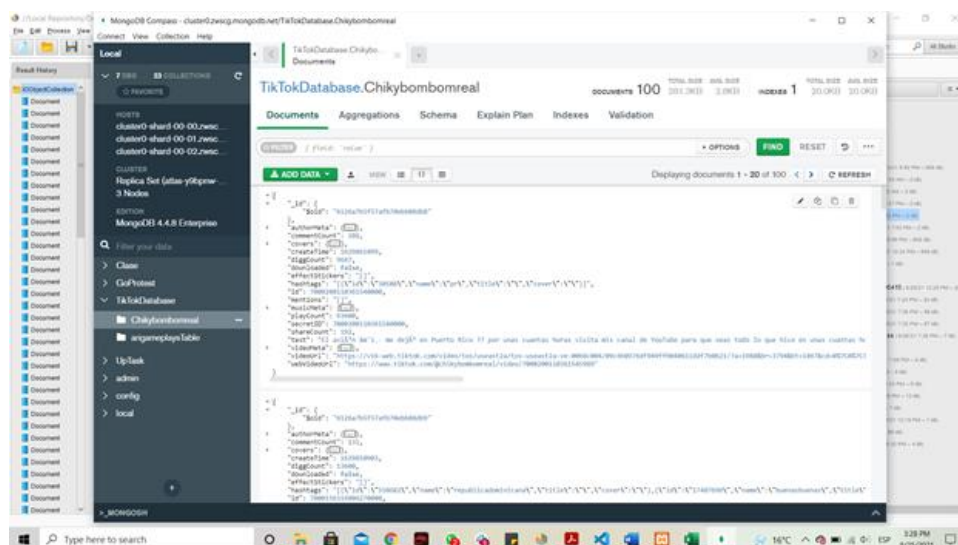
- En el operador Write Mongo se debe configurar la colección donde se van a almacenar los datos.



Después de correr el programa, se puede apreciar los datos en el cluster de MongoDB Atlas.

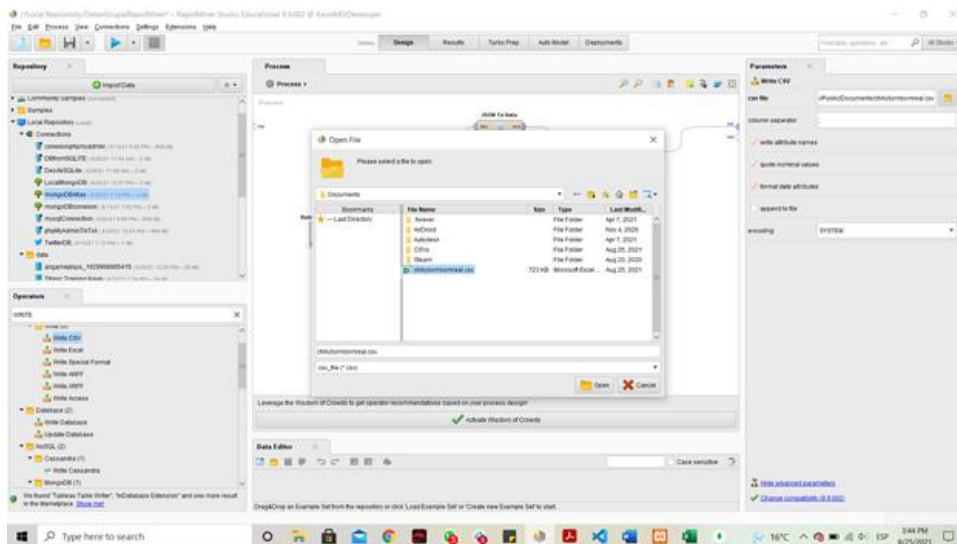


Datos de Chikybombomreal en MongoDB Atlas

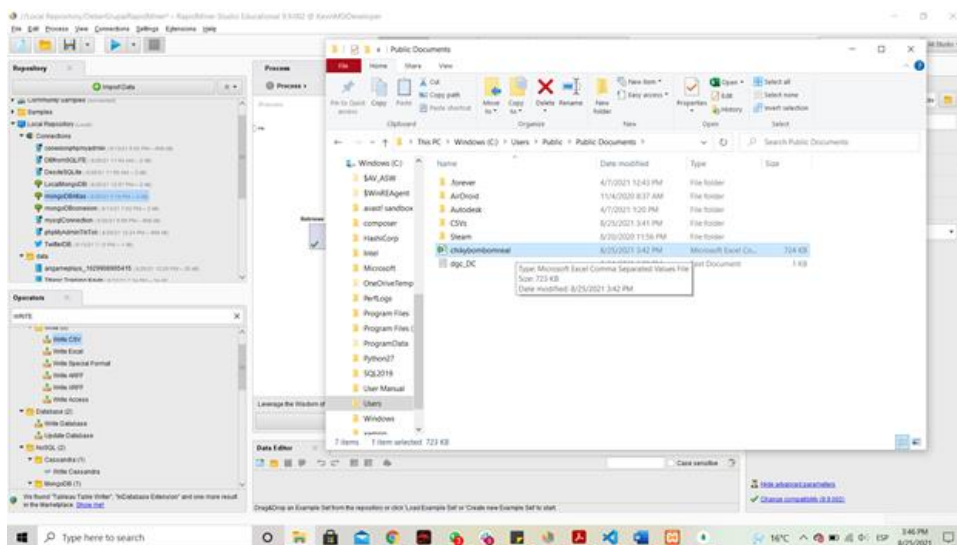


Verificación en MongoDB Atlas

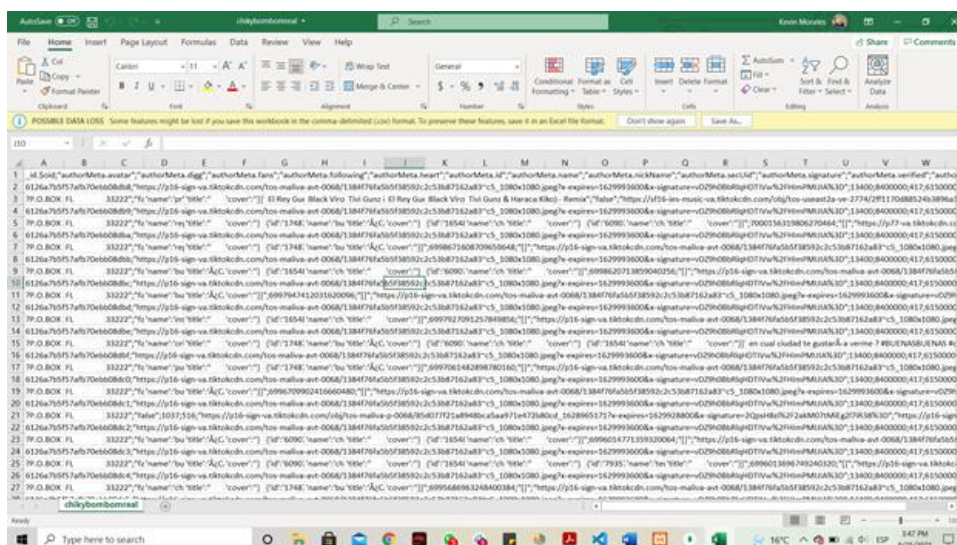




Evidencia del archivo CSV en los documentos públicos de mi PC.



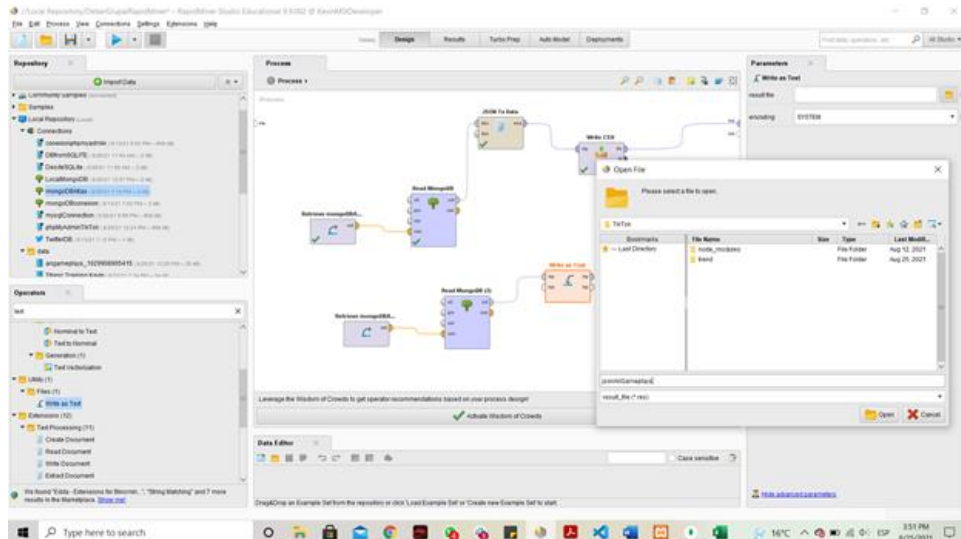
Archivo abierto en Excel, se procede a realizar lo mismo con la otra colección.



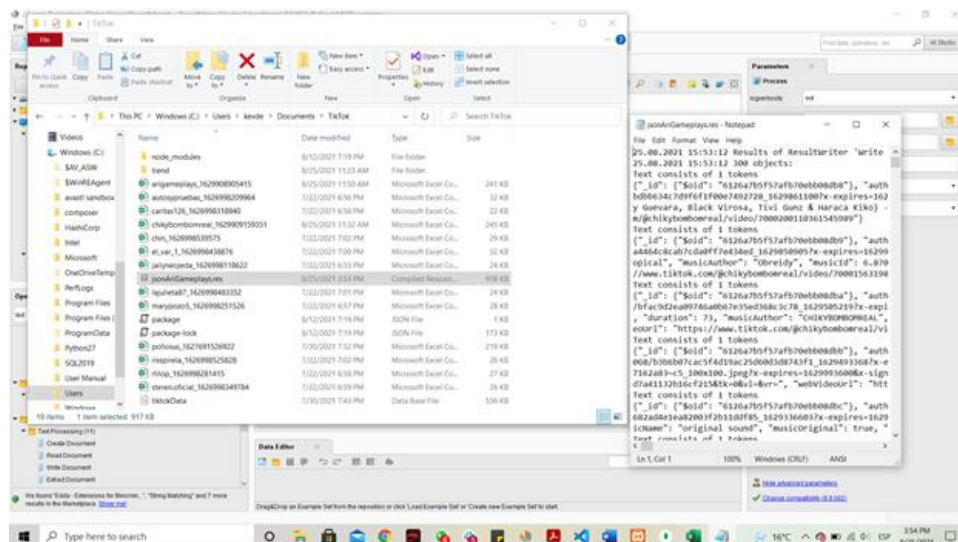
## • Parte 10



1. Para grabar un JSON basta con escribir el archivo en un documento de texto. Con el operador Write Text, escojo un lugar donde almacenar el archivo que proviene de MongoDB Atlas.



- Después, podemos verificar que ahora se tiene un archivo en donde se puede apreciar los objetos JSON de la base de dato proveniente de MongoDB Atlas.



En Visual Studio es posible ver los objetos JSON con mayor claridad:



