

Evaluación

# Bases de datos NoSQL

---

Ana María Forero Aldana

Máster Data Science, Big Data & Business Analytics 2023-2024

(Clase 1)

## Introducción

Siguiendo las indicaciones para desarrollar la tarea del módulo, se ha elegido el dataset público de la **Tate Collection**, la Galería Nacional de arte británico y moderno de Inglaterra, descargado de la plataforma *Opendatasoft* que alberga bases de datos de libre acceso para desarrolladores y programadores para potenciar sus proyectos de ciencia de datos.

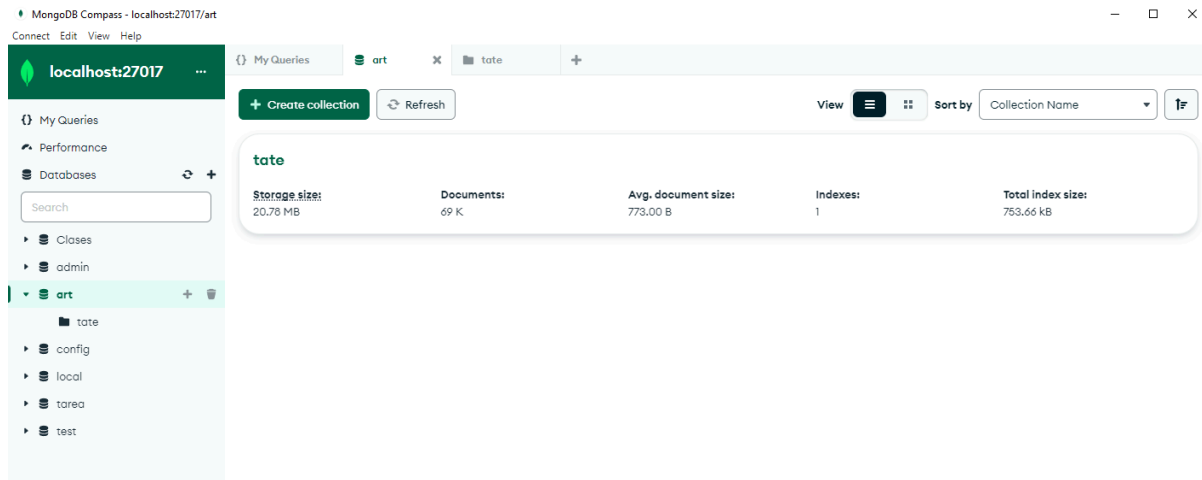
Este dataset contiene información de un total de 69.201 documentos que contienen datos de obras artísticas distribuidos en los siguientes campos:

- Artist
- Title
- Datetext
- Medium (soporte o técnica artística)
- CreditLine (un campo que nos da información sobre la procedencia de la obra, como si fue donada, comprada o adquirida)
- Year (de creación o ejecución de la obra)
- acquisitionYear (año de adquisición)
- dimensions
- Width (anchura)
- Height (altura)
- Depth (profundidad)
- units (unidades)
- Inscription
- thumbnailCopyright (Derechos de autor)
- URL
- Id (de la base de datos)
- Accession\_number(número de acceso de la obra en la colección)
- Artist\_id (identificador del artista)
- Artist\_role (rol del artista)

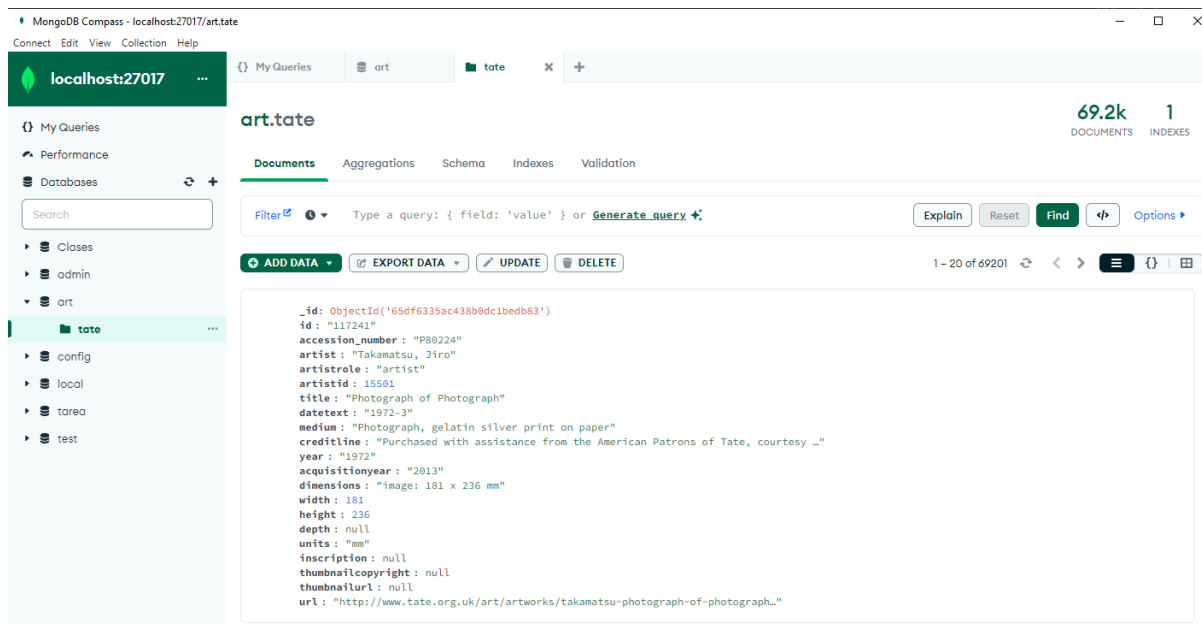
De cara a desarrollar la tarea será interesante averiguar datos tales como artista con más obras en la colección o técnica más representada, a través de la aplicación de queries en Mongo DB mediante el cliente NoSQL Booster.

## Cargar / importar dataset.

El dataset fue descargado de la plataforma de datos pública en formato JSON y cargada en la plataforma MongoDB Compass en la conexión local.



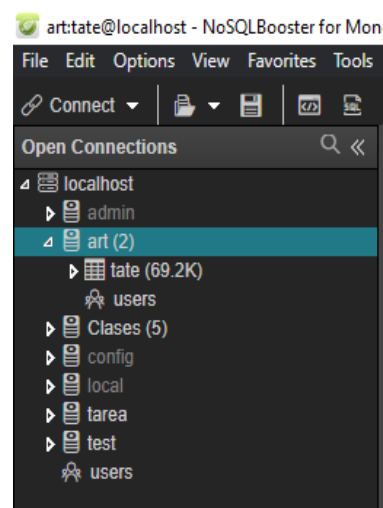
Al cargarla, fue llamada *art*, y la colección llamada *tate* (como su nombre original, Tate Collection).



Una vez cargada aquí, y tras conectar nuestra conexión local de MongoDB Compass con el cliente a través del cual vamos a ejecutar las queries, NoSQL Booster, refrescamos los conjuntos de datos de nuestra conexión local, actualizando así la base de datos recién ingresada, art.

## Ejercicios sobre inserción, actualización, proyección y filtrado.

Los directores de la Colección han adquirido la obra **La Toilette** de Henri Toulouse-Lautrec para aumentar así su catálogo de obras del artista francés, y quieren añadir los datos de la pintura a su dataset de obras de la colección.



Primeramente, queremos saber cuáles son las obras de Toulouse-Lautrec que almacena la Tate Collection. Para ello lanzamos la siguiente query, en la que filtramos por el campo 'artist' poniendo el nombre del artista, y creando una variable llamada *proyección* en la que le pedimos los campos que queremos que nos muestre. Finalmente, lo buscamos con la operación *find*.

```
var query = {'artist': 'Toulouse-Lautrec, Henri de'}
var proyeccion= {'artist': 1, 'title':1, 'year': 1, 'acquisitionyear': 1, 'medium':1, '_id':0 }
db.tate.find (query, proyeccion)
```

	artist	title	medium	year	acquisitionyear
1	Toulouse-Lautrec, Henri de	The Two Friends	Oil paint on board	1894	1940
2	Toulouse-Lautrec, Henri de	Side-saddle	Oil paint and gouache on board	1899	1983
3	Toulouse-Lautrec, Henri de	Emile Bernard	Oil paint on canvas	1885	1961

Vemos que tenemos tres obras del artista. Aplicaremos otra query de inserción en la que añadiremos los datos de la nueva obra adquirida. Esta vez creamos una nueva variable en la que completamos los campos con los datos de la nueva obra, que más adelante con la función *insertOne* añadirá los nuevos datos. Filtramos por el artista y los campos que queremos que muestre el resultado y buscamos la consulta con el *find*.

```
var nueva_obra = {'artist': 'Toulouse-Lautrec, Henri de', 'title': 'La toilette', 'medium': 'oil on canvas', 'year': '1896', 'acquisitionyear': '2024'}
var query = {'artist': 'Toulouse-Lautrec, Henri de'}
var proyeccion= {'artist': 1, 'title':1, 'year': 1, 'acquisitionyear': 1, 'medium':1, '_id':0 }
db.tate.insertOne(nueva_obra)
```

**db.tate.find(query, proyeccion)**

	artist ↕	title ↕	medium ↕	year ↕	acquisitionyear ↕
1	Toulouse-Lautrec, Henri de	The Two Friends	Oil paint on board	1894	1940
2	Toulouse-Lautrec, Henri de	Side-saddle	Oil paint and gouache on board	1899	1983
3	Toulouse-Lautrec, Henri de	Emile Bernard	Oil paint on canvas	1885	1961
4	Toulouse-Lautrec, Henri de	La toilette	oil on canvas	1896	2024

Y como se puede ver, se ha añadido la nueva obra a la base de datos de la Tate Collection.

Los gestores culturales de la Tate Collection quieren añadir un campo a su dataset que clasifique las obras producidas después de 1960 con la etiqueta de 'Arte contemporáneo'. Para ello se utiliza el operador de agregación *match* que filtra por el año indicado, y el operador *addFields*, que agrega o añade el valor del nuevo campo.

```
var etapa1 = { $match: { year: { $gte: '1960' } } }
var etapa2 = { $addFields: { 'Estilo': 'Arte Contemporáneo' } }
var proyeccion = {$project: {'artist': 1, 'title':1, 'year': 1, 'acquisitionyear': 1, 'medium':1, '_id':0, 'Estilo':1}}
var etapas = [etapa1, etapa2, proyeccion]
db.tate.aggregate(etapas).limit(10)
```

	artist ↕	title ↕	medium ↕	year ↕	acquisitionyear ↕	Estilo ↕
1	Takamatsu, Jiro	Photograph of Photograph	Photograph, gelatin silver print on paper	1972	2013	Arte Contemporáneo
2	Winters, Terry	Untitled	Graphite, chalk and watercolour on paper	1986	1987	Arte Contemporáneo
3	Solakov, Nedko	A Life (Black and White)	Performance, 2 people	1998	2009	Arte Contemporáneo
4	Le Brun, Christopher, PRA	[no title]	Etching on paper	1990	1991	Arte Contemporáneo
5	Christie, John	Banners No. 1	Screenprint on paper	1980	1981	Arte Contemporáneo
6	Bird, John	I.I.	Lithograph on paper	1973	1977	Arte Contemporáneo
7	Kapoor, Anish	Untitled 3	Intaglio print on paper	1988	1989	Arte Contemporáneo
8	Hockney, David	White Lines Dancing in Printing Ink	Lithograph on paper	1991	2004	Arte Contemporáneo
9	Newman, John	Second Thoughts III	Linocut on paper	1995	2004	Arte Contemporáneo
10	Stella, Frank	No Smoking (Small)	Enamel on steel	1998	2004	Arte Contemporáneo

## Ejercicios sobre pipeline de agregación.

La Tate Collection nos pide un informe para presentar a inversionistas sobre datos básicos de la Colección. Por ejemplo, nos pide que presentemos el artista del que más obras posee la Colección. Para ello utilizamos nuevos operadores. En un primer lugar, creamos una variable en la que utilizamos `group` para agrupar los documentos según una condición específica, que en este caso es el campo `'artist'`, y realizamos la agrupación a través del nuevo campo `'totalObras'` que a través del operador acumulativo `$sum`, hace una suma o conteo de cuántos documentos hay de cada grupo (definido antes por artista). En otra variable llamada etapa 2 utilizamos `$sort` para ordenar las obras de manera descendente, y en una tercera variable le pedimos que limite el resultado a 1, que sería el primero.

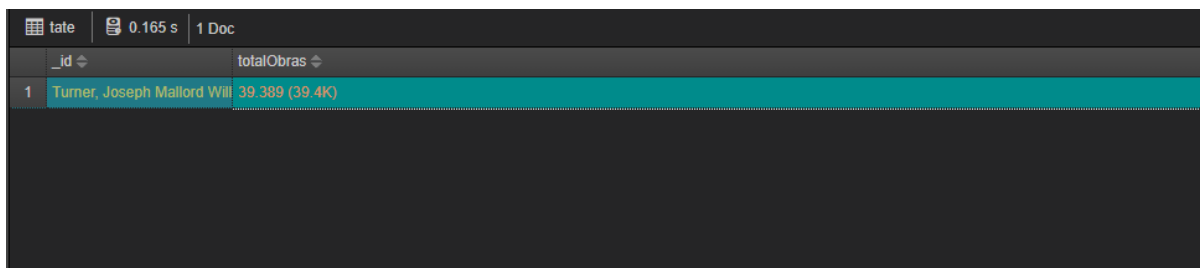
```
var etapa1 = { $group: { _id: "$artist", totalObras: { $sum: 1 } } };
```

```
var etapa2 = { $sort: { totalObras: -1 } };
```

```
var etapa3 = { $limit: 1 };
```

```
var etapas = [etapa1, etapa2, etapa3];
```

```
db.tate.aggregate(etapas);
```



The screenshot shows a MongoDB interface with a query result. The top bar indicates the database is 'tate', the query took 0.165 seconds, and there is 1 document. The result is a table with two columns: '\_id' and 'totalObras'. The first row shows 'Turner, Joseph Mallord Will' as the artist and '39,389 (39.4K)' as the total number of works.

	_id	totalObras
1	Turner, Joseph Mallord Will	39,389 (39.4K)

Esta query nos lanza como resultado que el artista del que más obras posee la Colección es **J.M.W.Turner**, uno de los pintores románticos más importantes del arte inglés, con un total de 39.389 obras.

Dentro de las personas de interés en la Colección se encuentran gestores culturales que buscan constantemente artistas de los que hacer exposiciones temporales en sus lugares de origen. En concreto les gustaría conocer la producción de Turner con una técnica artística concreta, que es el grafito sobre papel, pues es una técnica muy usada por muchos artistas de diferentes épocas para perfeccionar sus técnicas de dibujo. En concreto, nos piden conocer cuántas obras de Turner tienen esta técnica.

Lanzamos una query en la que filtramos por artista y técnica (o medio), y le pedimos que nos proyecte el medio.

```
var query = {'artist': 'Turner, Joseph Mallord William', 'medium': 'Graphite on paper'}
var proyeccion = {'medium': 'Graphite on paper'}
db.tate.find (query, proyeccion)
```

	_id	medium
1	65df6335ac438b0dc1be	Graphite on paper
2	65df6335ac438b0dc1be	Graphite on paper
3	65df6335ac438b0dc1be	Graphite on paper
4	65df6335ac438b0dc1be	Graphite on paper
5	65df6335ac438b0dc1be	Graphite on paper
6	65df6335ac438b0dc1be	Graphite on paper
7	65df6335ac438b0dc1be	Graphite on paper
8	65df6335ac438b0dc1be	Graphite on paper
9	65df6335ac438b0dc1be	Graphite on paper
10	65df6335ac438b0dc1be	Graphite on paper
11	65df6335ac438b0dc1be	Graphite on paper
12	65df6335ac438b0dc1be	Graphite on paper
13	65df6335ac438b0dc1be	Graphite on paper
14	65df6335ac438b0dc1be	Graphite on paper
15	65df6335ac438b0dc1be	Graphite on paper
16	65df6335ac438b0dc1be	Graphite on paper
17	65df6335ac438b0dc1be	Graphite on paper
18	65df6335ac438b0dc1be	Graphite on paper
19	65df6335ac438b0dc1be	Graphite on paper
20	65df6335ac438b0dc1be	Graphite on paper
21	65df6335ac438b0dc1be	Graphite on paper
22	65df6335ac438b0dc1be	Graphite on paper
23	65df6335ac438b0dc1be	Graphite on paper
24	65df6335ac438b0dc1be	Graphite on paper
25	65df6335ac438b0dc1be	Graphite on paper

El resultado nos devuelve más de 24 mil obras del artista sólo con esta técnica. El artista se destaca por obras de temática paisajística, sin embargo más de la mitad de la colección son grabados sobre papel, lo que puede dejar ver que la Colección posee una innumerable muestra de obras en las que iba perfeccionando sus técnicas ilustrativas.

Pensando en Turner como artista romántico, las personas de interés en la Tate Collection querrían saber qué obras ejecutadas antes de 1850 posee la Colección. Para ello, se filtra por el campo 'year' y se muestran los campos deseados.

```
var query = {'year': { '$lt': '1850' }}
var proyeccion= {'artist': 1, 'title':1, 'year': 1, 'acquisitionyear': 1, 'medium':1, '_id':0 }
```

db.tate.find (query, proyeccion)

	artist	title	medium	year	acquisitionyear
1	Blake, William	First Book of Urizen pl	Etching with paint, watercolour and ink	1796	2010
2	Turner, Joseph Mallord William	Mountain Sketch; Skeb	Graphite on paper	1844	1856
3	Turner, Joseph Mallord William	Mountain Scenery and	Graphite on paper	1833	1856
4	Turner, Joseph Mallord William	Castle on Rock	Graphite on paper	1830	1856
5	Turner, Joseph Mallord William	?Mountain Pass	Graphite and watercolour on paper	1843	1856
6	Turner, Joseph Mallord William	Knaresborough, Yorkst	Line engraving on paper	1828	1986
7	Turner, Joseph Mallord William	The Lake of Geneva	Line engraving on paper	1830	1986
8	Turner, Joseph Mallord William	Château de Nantes	Intaglio print on paper	1833	1989
9	Turner, Joseph Mallord William	Breakers on a Flat Bea	Oil paint on canvas	1835	1856
10	Turner, Joseph Mallord William	Lecture Diagram: Anan	Graphite and watercolour on paper	1817	1856
11	Turner, Joseph Mallord William	Views of a Town. Newt	Graphite on paper	1825	1856

El resultado nos devuelve que casi la mitad de las obras que posee la Tate Collection fueron ejecutadas antes de 1850, lo que demuestra un gran catálogo de obras que pueden ser de muchas corrientes que van desde el Romanticismo hasta el impresionismo.

Por otro lado, queremos saber las obras que tiene la Tate Collection cuyo año de ejecución fue entre 1850 y 1900. Para ello utilizamos el operador lógico *and* que devuelve sólo las condiciones indicadas.

```
var query1= { 'year': { $gte: '1850' } }
var query2= { 'year': { $lte: '1900' } }
var logic= { $and: [query1, query2] }
var proyeccion= {'artist': 1, 'title':1, 'year': 1, 'acquisitionyear': 1, 'medium':1, '_id':0 }
db.tate.find ( logic, proyeccion )
```



tate 0.038 s 1509 Docs					
	artist	title	medium	year	acquisitionyear
1	Sandys, Frederick	Great Yarmouth and B	Chalk on paper	1871	1989
2	Ridley, Matthew White	The Pool of London	Oil paint on canvas	1862	1919
3	Inchbold, John William	Tintagel	Graphite and watercolour on paper	1861	1997
4	Fry, Roger	Landscape with Sheph	Oil paint on canvas	1891	1973
5	Maitland, Paul	Barges, Chelsea River	Oil paint on wood	1885	1948
6	Constable, John	Summer, Afternoon - A	Mezzotint on paper	1855	1985
7	Burne-Jones, Sir Edward Coley, Bt	Study of Ezekiel's Han	Graphite on paper	1860	1927
8	Conder, Charles	Fan: The Romantic Ex	Watercolour on silk	1899	1917
9	Brown, Ford Madox	Study of the Tow-Path	Graphite on paper	1890	1898
10	Lear, Edward	Porto Venere	Ink and watercolour on paper	1860	1910
11	Toulouse-Lautrec, Henri de	The Two Friends	Oil paint on board	1894	1940

Por el contrario, se puede apreciar el contraste en comparación con los datos de la consulta anterior, pues la Colección posee solo 1509 obras ejecutadas entre 1850 y 1900, época de grandes cambios a nivel artístico, también en parte gracias al surgimiento de la fotografía, en la que los artistas se enfrentaron a la necesidad de producción de nuevas técnicas y corrientes más abstractas, como vendría a ser más adelante, con estilos como el vanguardismo.

Los gestores culturales que visitan la Colección están interesados en aquellas obras sin nombre, es decir, aquellas que no tienen título ni año de ejecución, realizadas con una técnica concreta que es la tinta sobre papel. Realizamos una query que nos de estos datos.

```
var query1 = {'year':{$type:'null'}}
var query2 = {'title':['title not known']}
var query3 = {'acquisitionyear': {$gte: '1850', $lte: '1900'}}
var query4 = {'medium':'Ink on paper'}
var logic = {$and: [query1, query2, query3, query4]}
var proyeccion= {'artist': 1, 'title':1, 'year': 1, 'acquisitionyear': 1, 'medium':1, '_id':0 }
db.tate.find ( logic, proyeccion )
```

tate

0.054 s







352 Docs

	artist	title	medium	year	acquisitionyear
1	Jones, George	[title not known]	Ink on paper	null	1888
2	Jones, George	[title not known]	Ink on paper	null	1888
3	Jones, George	[title not known]	Ink on paper	null	1888
4	Jones, George	[title not known]	Ink on paper	null	1888
5	Jones, George	[title not known]	Ink on paper	null	1888
6	Jones, George	[title not known]	Ink on paper	null	1888
7	Jones, George	[title not known]	Ink on paper	null	1888
8	Jones, George	[title not known]	Ink on paper	null	1888
9	Jones, George	[title not known]	Ink on paper	null	1888
10	Jones, George	[title not known]	Ink on paper	null	1888
11	Jones, George	[title not known]	Ink on paper	null	1888

El resultado nos lanza 352 obras con estas características, ejecutadas por el artista George Jones, pintor británico de la primera mitad del siglo XIX que destacó por sus dibujos de temática militar. Este resultado puede ser interesante si se quiere rescatar su obra para una exposición sobre el artista o sus dibujos con la misma técnica.

Viendo la relevancia que tienen las obras de tinta o grafito sobre papel, los gestores se preguntan cuál es la técnica más representada en la Colección.

```
var etapa1 = { $group: { _id: "$medium", count: { $sum: 1 }, works: { $push: "$title" } } }
var etapa2 = { $sort: { count: -1 } }
var etapa3 = { $limit: 1 }
var etapas = [etapa1, etapa2, etapa3]
db.tate.aggregate(etapas)
```

 tate	 0.521 s	1 Doc	
	 _id	 count	 works
1	Graphite on paper	26.167 (26.2k)	 Array[26167]

La técnica más utilizada en las obras que contiene la Tate Collection es la de grafito sobre papel.



## Conclusiones

El estudio del dataset de la Tate Collection puede ayudar a la toma de decisiones al responder preguntas concretas como conocer qué obras hay de un artista en concreto, añadir una etiqueta de estilo artístico a obras que cumplen determinadas condiciones, conocer la técnica artística más utilizada en las obras de la colección, o el artista del que más obras se tiene en la misma. Esta información, que podemos extraer aplicando determinadas queries concretas a través de Mongo DB, responde a esa demanda cada vez más exigente de información más precisa que conduzca a las organizaciones culturales a continuar con su labor divulgativa del arte y la cultura.



## Recursos

Tate Collection → [Collection | Tate](#)

Opendatasoft → [Explore — Opendatasoft](#)

