01MBID Fundamentos de la tecnología Big Data

Universidad Internacional de Valencia

Sesión 3

Tema 2 + Actividad

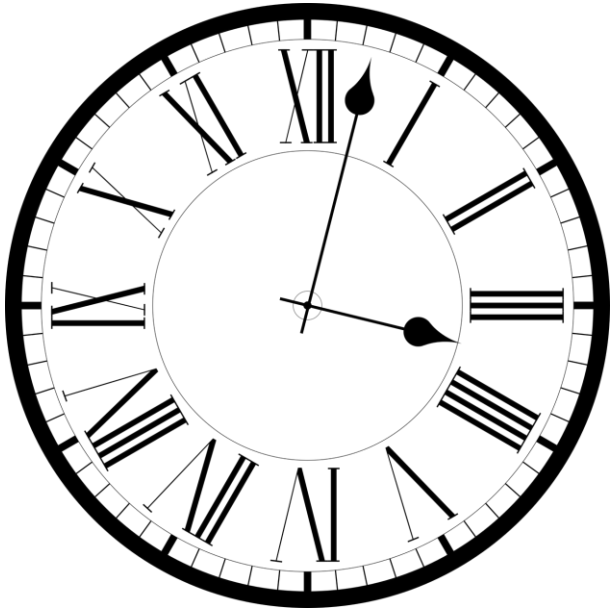
De:
Planeta Formación y Universidades

Agenda

- **Dudas**
- **Tema 2 2de2: Fuentes de datos en entornos Big data**
- **Actividad**

Agenda

- **Dudas**
- **Tema 2 2de2: Fuentes de datos en entornos Big data**
- **Actividad**



Agenda

- Dudas
- **Tema 2 2de2: Fuentes de datos en entornos Big data**
- **Actividad**

Fuentes de datos en entornos Big data

1) ¿Qué es una fuentes de datos?
2) Diferencias respecto a las tecnologías de datos tradicionales
3) **Tipos de datos y flujo de datos**
4) **Calidad de Datos**
5) **Las V's del Big Data**

Tipos de datos y flujos de datos

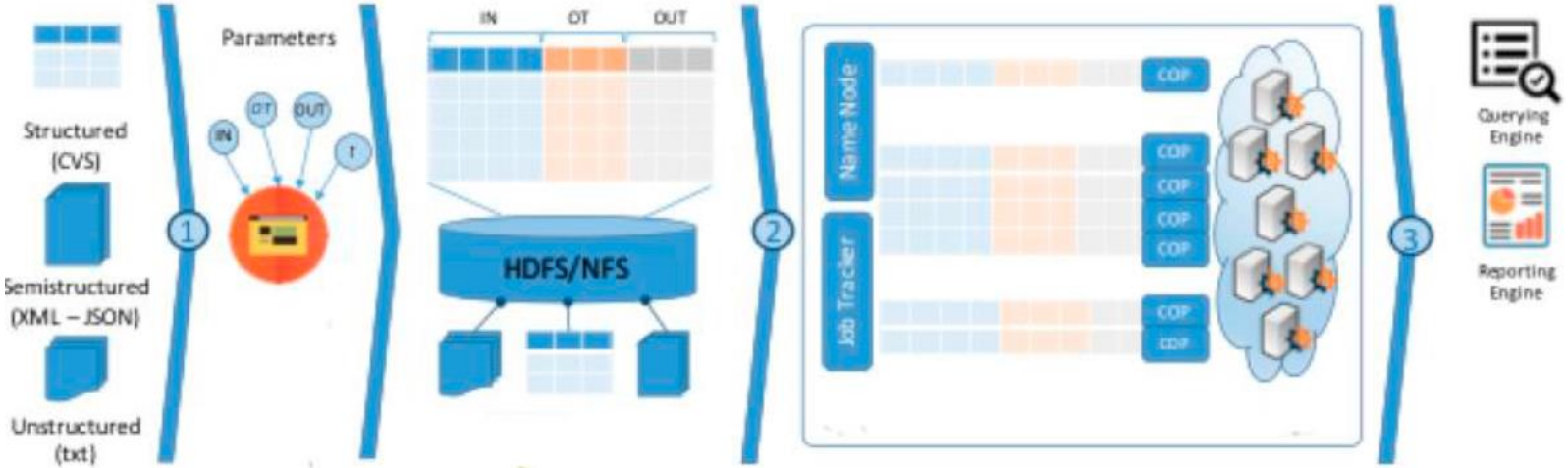
- Estructurados
  - formato + finitos
  - ejemplo, expediente académico

- No Estructurados
  - sin formato + indeterminado + herramientas especificas
  - ejemplo, video

- Semiestructurados
  - marcadores o organización interna
  - ejemplo, XML

Tipos de datos y flujos de datos

Ejemplo:



Fuentes de datos en entornos Big data

1) ¿Qué es una fuentes de datos?
2) Diferencias respecto a las tecnologías de datos tradicionales
3) Tipos de datos y flujo de datos
4) **Calidad de Datos**
5) **Las V's del Big Data**

Calidad de datos

ISO 25012:2008

Grado en que las características de los datos satisfacen las necesidades explícitas e implícitas cuando se usan bajo condiciones específicas.

"Calidad adecuación al uso"

Importancia en Big Data

- Objetivo: definir para qué y por qué se quieren los datos. ¿Qué **valor** me aporta?

- **Dimensiones de Calidad**: disponibilidad, usabilidad, confiabilidad, pertinencia y presentación.

Importancia en Big Data

Condición previa para el análisis y el uso de Big Data y para garantizar el valor de esos datos.

Calidad de datos

- Disponibilidad
- Usabilidad
- Confiabilidad
- Pertinencia
- Presentación

Disponibilidad

Accesibilidad:

- Interfaz de acceso a datos
- Los datos públicos o fáciles de adquirir

Oportunidad:

- Rango de tiempo, para que sean útiles
- Actualizados “regularmente”
- Intervalo de tiempo entre la recopilación, procesamiento y uso cumple los requisitos

Usabilidad

Credibilidad:

- Fuente de los datos: organizaciones especializadas de un país, campo o industria
- Expertos o especialistas auditan regularmente y comprueban la exactitud de los datos
- Los datos existen en el rango de valores conocidos o “aceptables”

Confiabilidad

Exactitud:

- Los datos proporcionados son precisos
- La representación de datos (o valor) refleja bien el estado real de la información de origen
- La representación de información (datos) no causara ambigüedad

Consistencia:

- Después de procesar los datos, sus conceptos, dominios de valor y formatos coinciden con los originales
- Durante un cierto tiempo, los datos permanecen consistentes
- Durante un cierto tiempo, los datos permanecen verificables

Confiabilidad

Integridad:

- El formato de los datos es claro y cumple los criterios
- Los datos son consistentes con la integridad estructural
- Los datos son consistentes con la integridad del contenido

Completitud:

- Si una deficiencia de un componente afectara el uso de los datos para datos con componentes múltiples
- Si una deficiencia de un componente afectara la precisión y la integridad de los datos

Pertinencia

Conveniencia:

- Los datos recogidos no coinciden completamente con el tema, pero exponen un aspecto parcial del mismo.
- La mayoría de los conjuntos de datos recuperados están dentro de la temática de recuperación que los usuarios necesitan
- La temática de la información proporciona coincidencias con la temática de recuperación de los usuarios

Presentación

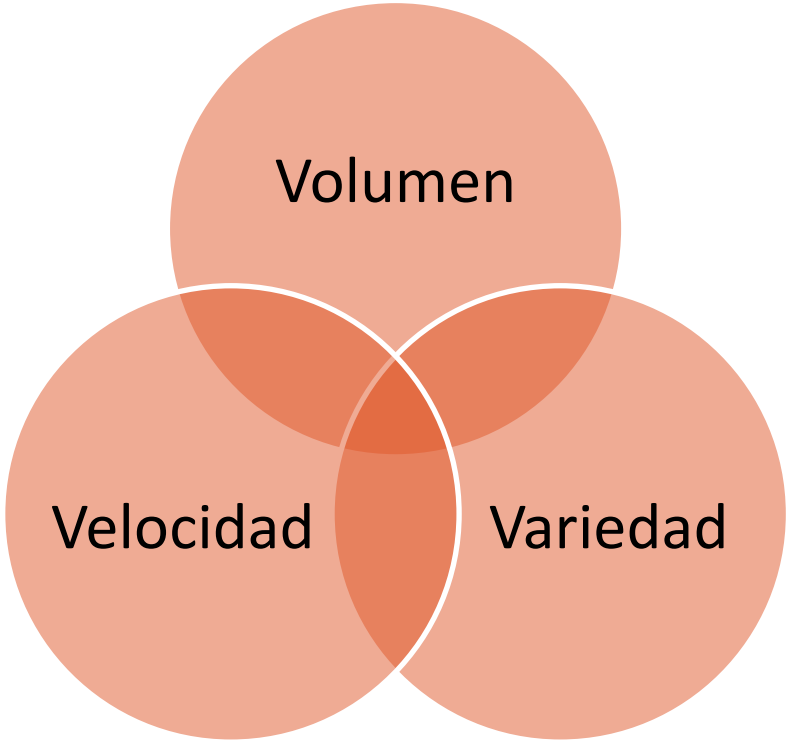
Legibilidad:

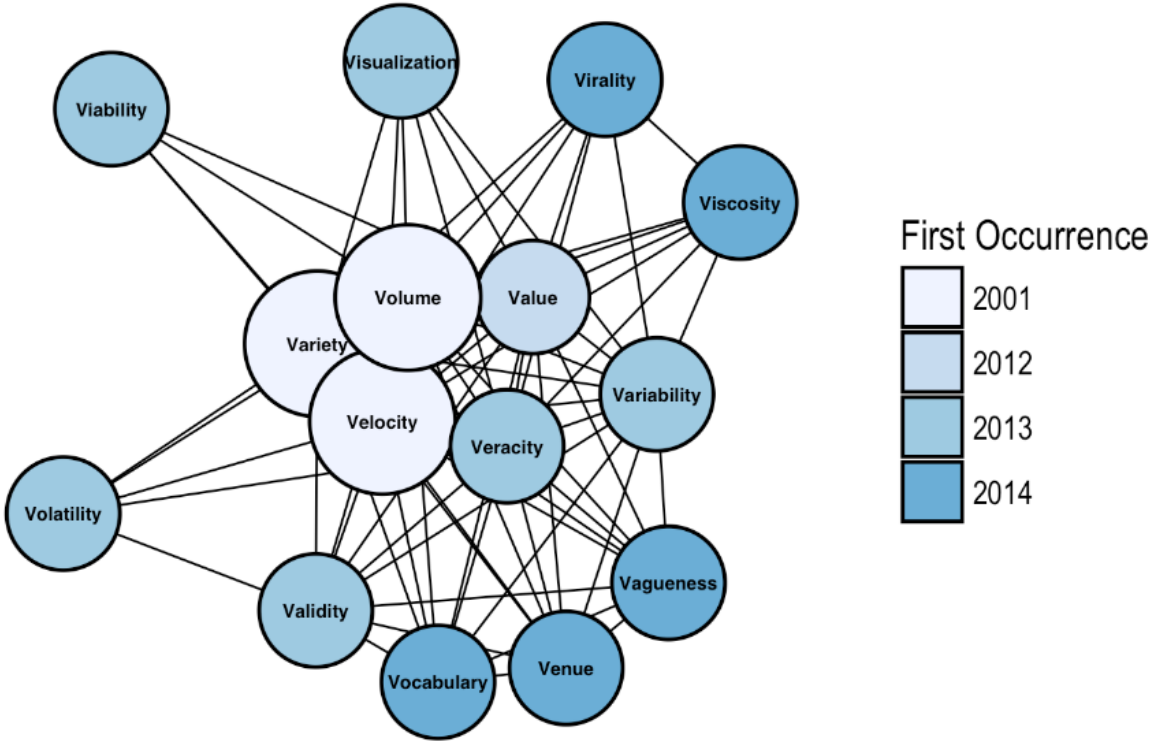
- Los datos (contenido, formato, etc.) son claros y comprensibles
- Es fácil juzgar que los datos facilitados satisfacen las necesidades
- La descripción de los datos, la clasificación y el contenido de codificación satisfacen la especificación y son fáciles de entender

Fuentes de datos en entornos Big data

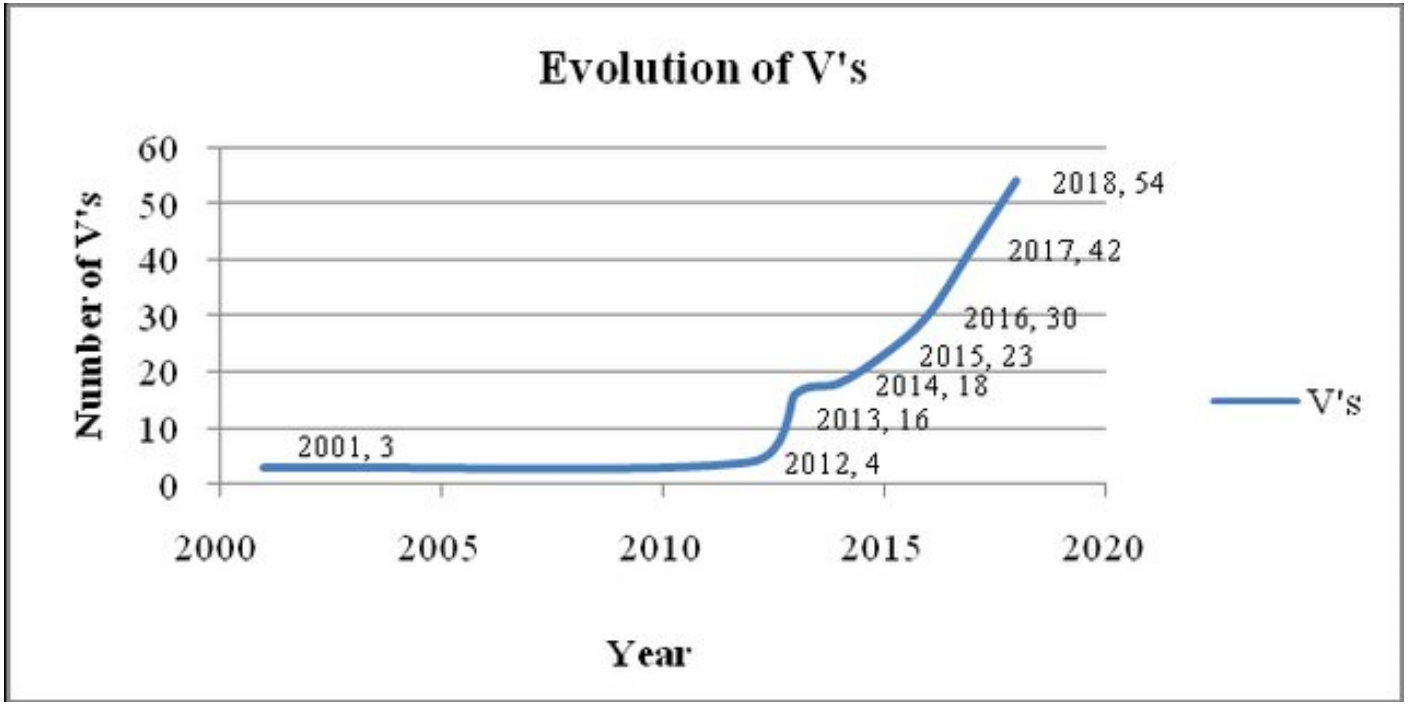
1) ¿Qué es una fuentes de datos?
2) Diferencias respecto a las tecnologías de datos tradicionales
3) Tipos de datos y flujo de datos
4) Calidad de Datos
5) **Las V's del Big Data**

Las V's del Big Data



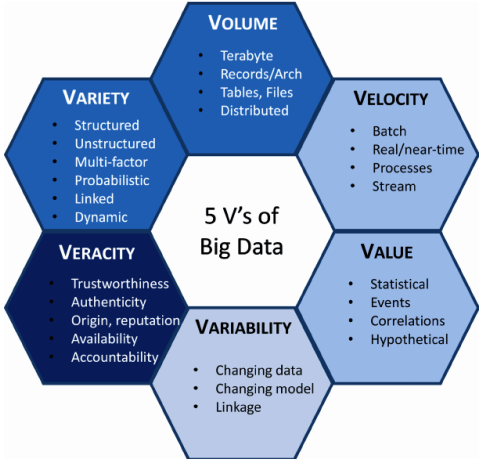


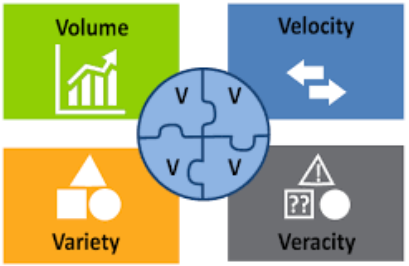
* https://www.elderresearch.com/blog/42-v-of-big-data

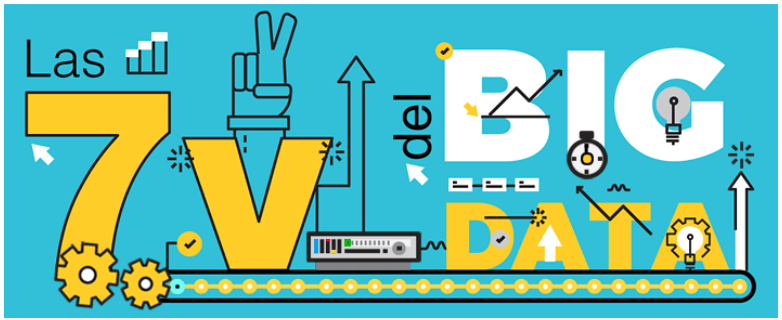


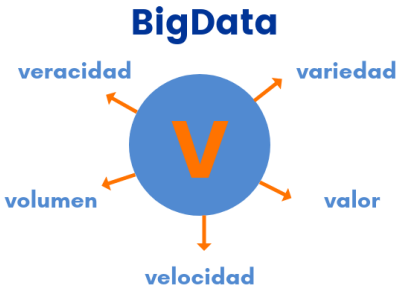
* https://www.researchgate.net/publication/335619326_Unlock_Different_V's_of_Big_Data_for_Analytics

Las V's del Big Data









Información sobre las V's

Infografía Big Data: las 7 V
http://www.iic.uam.es/innovacion/big-data-infografia-7-v/
https://www.youtube.com/watch?v=S81frJYbjcE

Big Data Technology with 8 V´s
https://www.m-brain.com/technology/

The 10 Vs of Big Data
https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx

Información sobre las V's

Las siete 'V' del Big Data
https://bbvaopen4u.com/es/actualidad/las-siete-v-del-big-data

The 42 V's of Big Data and Data Science
* https://www.elderresearch.com/blog/42-v-of-big-data
https://www.kdnuggets.com/2017/04/42-vs-big-data-data-science.html

Las 10 V´s del Big Data
https://www.grupo-novatech.com/las-10-vs-del-big-data/

How to Ensure the Validity, Veracity, and Volatility of Big Data
https://www.dummies.com/programming/big-data/engineering/how-to-ensure-the-validity-veracity-and-volatility-of-big-data/

Información sobre las V's

The Eight V's of Supercomputing and Big Data
https://www.nimbix.net/eight-vs-supercomputing-big-data

BIG DATA AND FIVE V'S CHARACTERISTICS
https://www.researchgate.net/publication/332230305_BIG_DATA_AND_FIVE_V'S_CHARACTERISTICS

Big Data: el futuro a través de los datos
http://www.upm.es/sfs/Rectorado/Gabinete%20del%20Rector/Revista%20UPM/NUMERO_31_VER.pdf

Los Big Data: Conceptos relacionados y algunas aplicaciones en pediatría
https://scielo.conicyt.cl/scielo.php?script=sci_arttext&pid=S0370-41062019000400376#B13

Información sobre las V's

Grandes datos, grandes desafíos para las ciencias sociales
http://www.scielo.org.mx/scielo.php?pid=S0188-25032018000200415&script=sci_arttext

The 23Vs of Big Data
https://hackernoon.com/the-23vs-of-big-data-c9146716e0cb

The five V's of big data
https://www.bbva.com/en/five-vs-big-data/

¿Cuáles son las 5 V's del Big Data?
https://www.iebschool.com/blog/5-vs-del-big-data/

Información sobre las V's

WHAT ARE THE 7 V'S OF BIG DATA?
https://www.yourtechdiet.com/blogs/7vs-big-data/

LAS 10 V'S DEL BIG DATA
https://www.datahack.es/blog/big-data/10-vs-del-big-data/

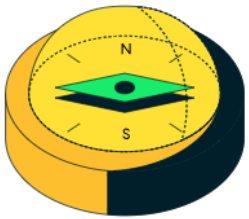
…

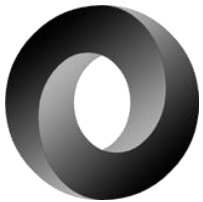
Agenda

- Dudas
- Tema 2 2de2: Fuentes de datos en entornos Big data
- **Actividad**

Actividad

MongoDB + Python + JSON



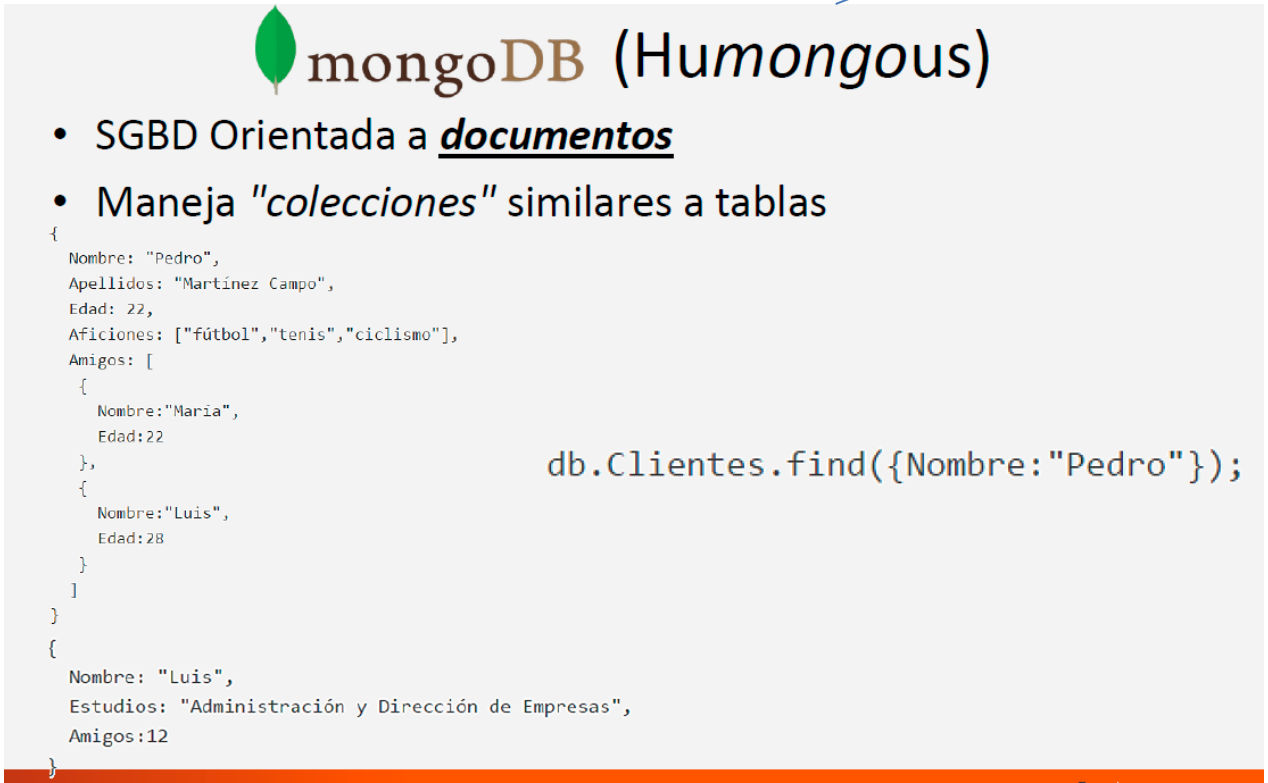


Conceptos clave

-MongoDB
-JSON

Muy grande

MongoDB

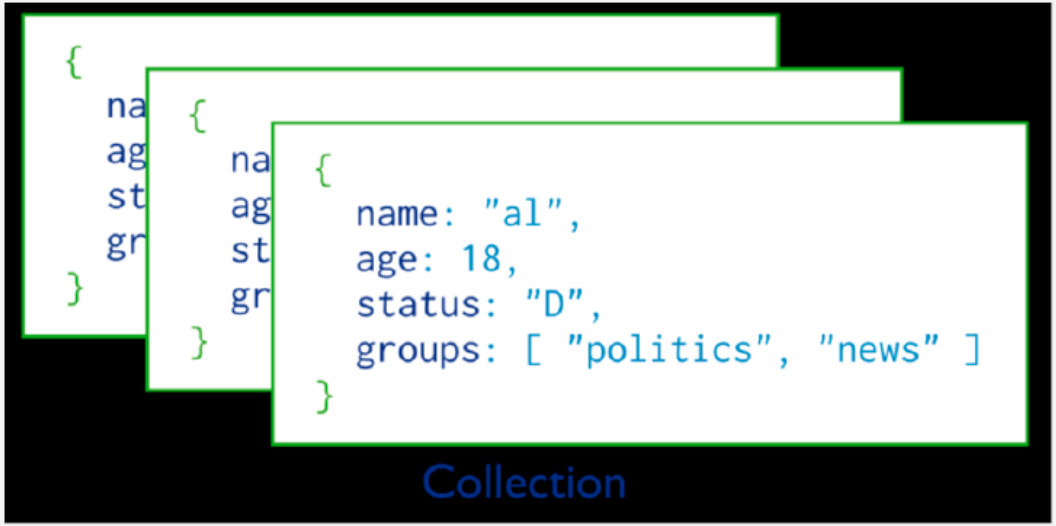


MongoDB

Documento

MongoDB

Colección



Conjunto de documentos no necesariamente con la misma estructura

MongoDB

Modelo Relacional	MongoDB
Tabla	Colección
registro	documento
columna	campo

https://www.mongodb.com/docs/manual/reference/sql-comparison/#sql-to-mongodb-mapping-chart

MongoDB

- MongoDB almacena internamente los datos en BSON, abreviatura de Binary-JSON.

- JSON es *Java Script Object Notation*, una notación estándar para el intercambio de datos.

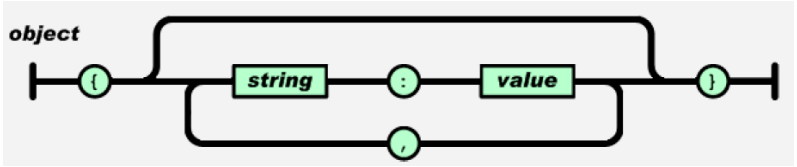
- JSON especifica una gramática mediante expresiones regulares.

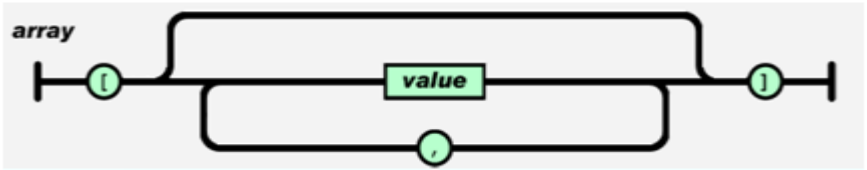
MongoDB
JSON

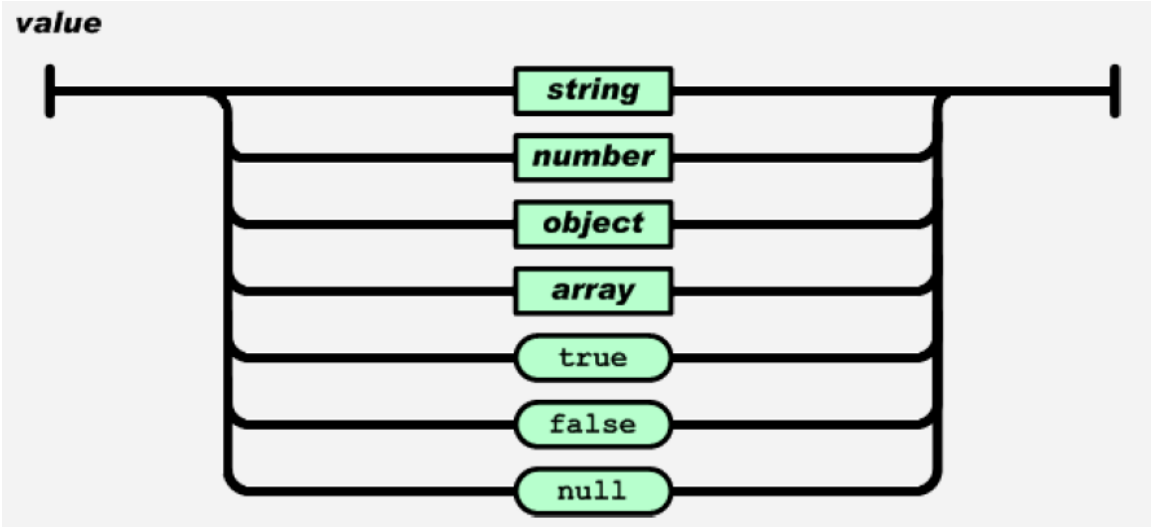
```
object::='{' members' }'*
members::=
      pair | pair ',' members
pair::= string : value
array::= '[' element ']'*
element::=
      value| value ',' element
```

```
value:=
    String |
    number |
    object |
    array |
    true |
    false |
    null
```

JSON


object
{
string
:
value
}
,


array
[
value
]
,


value
string
number
object
array
true
false
null

Diccionario Python = JSON

```
{"Space X":
 [
  {"VehiculosLanzamiento":
   [
    {"Nombre":"Falcon 1","CargaUtil":670},
    {"Nombre":"Falcon 9","CargaUtil":22800},
    {"Nombre":"Falcon Heavy","CargaUtil":63800}
   ]
  },
  {"Naves":
   [
    {"Nombre":"Dragon 1","Descripcion":"Cargo Dragon"},
    {"Nombre":"Dragon 2","Descripcion":"Crew Dragon"}
   ]
  }
 ]
}
```

```
{
  "Space X": [
    {
      "VehiculosLanzamiento": [
        {
          "Nombre": "Falcon 1",
          "CargaUtil": 670
        },
        {
          "Nombre": "Falcon 9",
          "CargaUtil": 22800
        },
        {
          "Nombre": "Falcon Heavy",
          "CargaUtil": 63800
        }
      ]
    },
    {
      "Naves": [
        {
          "Nombre": "Dragon 1",
          "Descripcion": "Cargo Dragon"
        },
        {
          "Nombre": "Dragon 2",
          "Descripcion": "Crew Dragon"
        }
      ]
    }
  ]
}
```

** https://jsonformatter.org/*

MongoDB - Twitter

```
{_id:POST_ID,
 title:TITLE_OF_POST,
 description:POST_DESCRIPTION,
 by:POST_BY,
 url:URL_OF_POST
 tags:[TAG1,TAG2,TAG3],
 likes:TOTAL_LIKES,
 comments:
   [
{user:'COMMENT_BY',message:TEXT,dateCreated:DATE_TIME,like:LIKES},
{user:'COMMENT_BY',message:TEXT,dateCreated:DATE_TIME,like:LIKES}
   ]
}
```

Los tags se integran como array.

Los comentarios se integran como array de subdocumentos

MongoDB Consultas + Actualizaciones

Equivalencias de SQL en MongoDB

https://www.mongodb.com/docs/manual/reference/sql-comparison/#sql-to-mongodb-mapping-chart

SQL Terms, Functions, and Concepts	MongoDB Aggregation Operators
WHERE	$match
GROUP BY	$group
HAVING	$match
SELECT	$project
ORDER BY	$sort
LIMIT	$limit
SUM()	$sum
COUNT()	$sum
join	No existe, se puede hacer algo similar con $unwind con los campos embebidos en un subdocumento.

MongoDB CRUD Operations

https://docs.mongodb.com/manual/crud/

- find
- count
- sort
- group
- sum
- unwind
- limit

MongoDB Pipeline

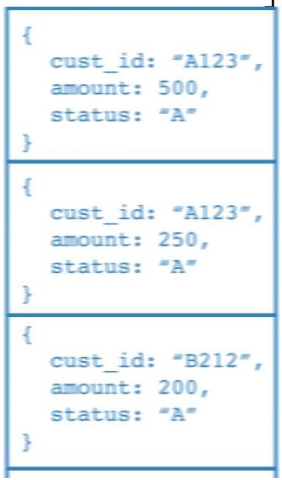
https://docs.mongodb.com/manual/core/aggregation-pipeline/

```
db.orders.aggregate( [
                    { $match: { status: “A” } },
                    { $group: { _id: “$cust_id”,total:  {$sum: “$amount” } } }
                    ])
```

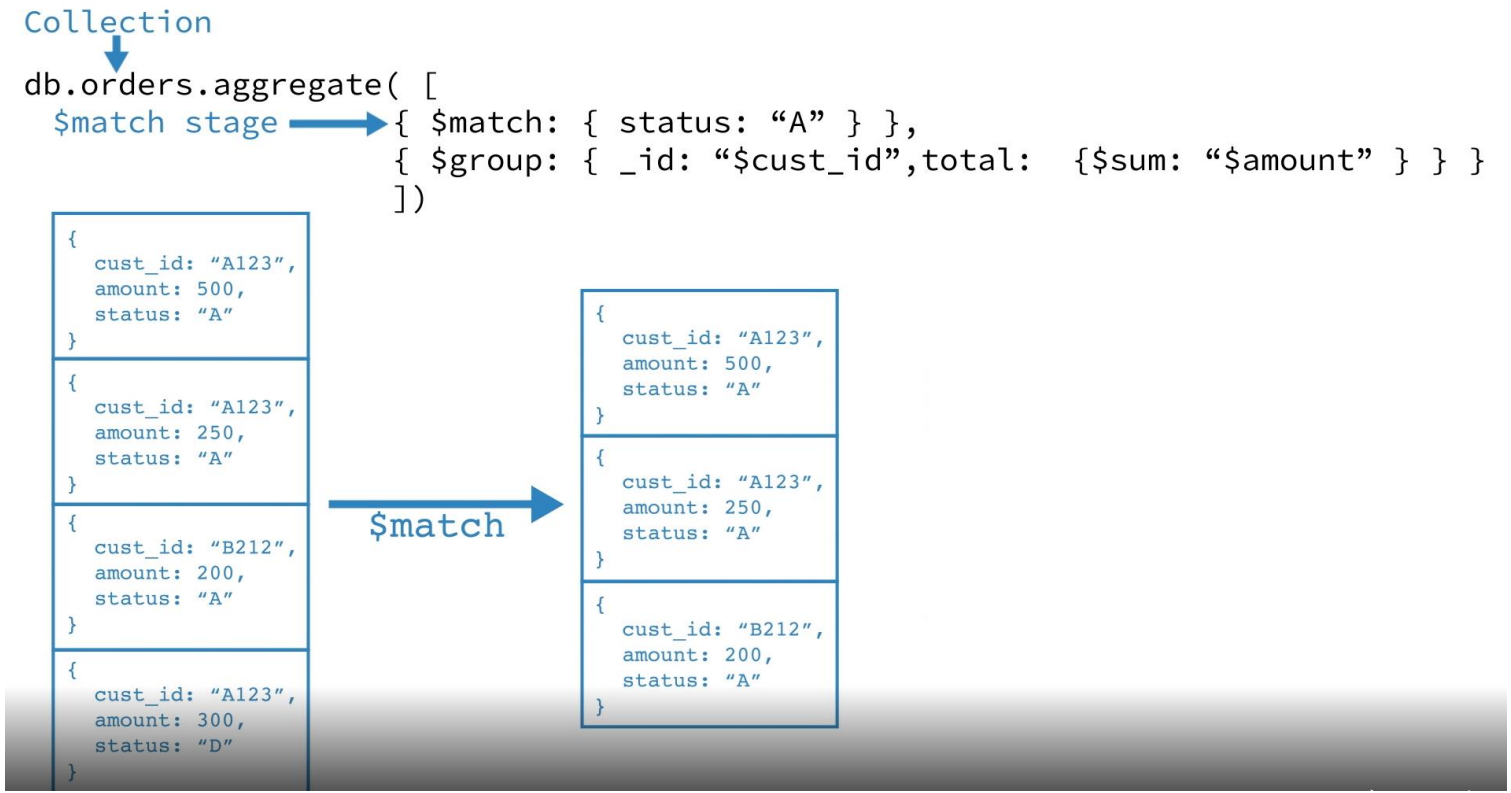
MongoDB Pipeline

Collection

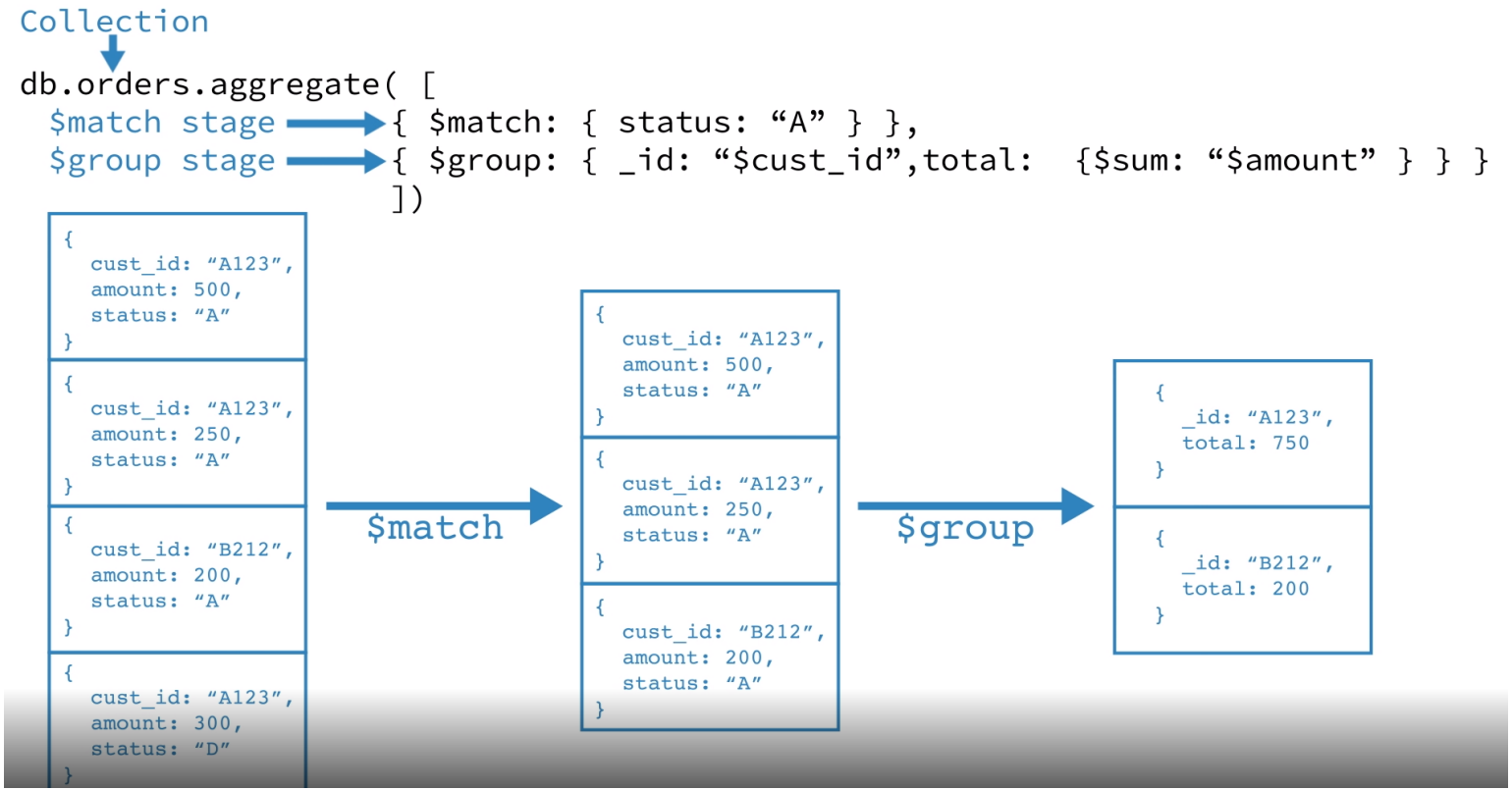
```
db.orders.aggregate( [
                    { $match: { status: “A” } },
                    { $group: { _id: “$cust_id”,total:  {$sum: “$amount” } } }
                    ])
```



MongoDB Pipeline



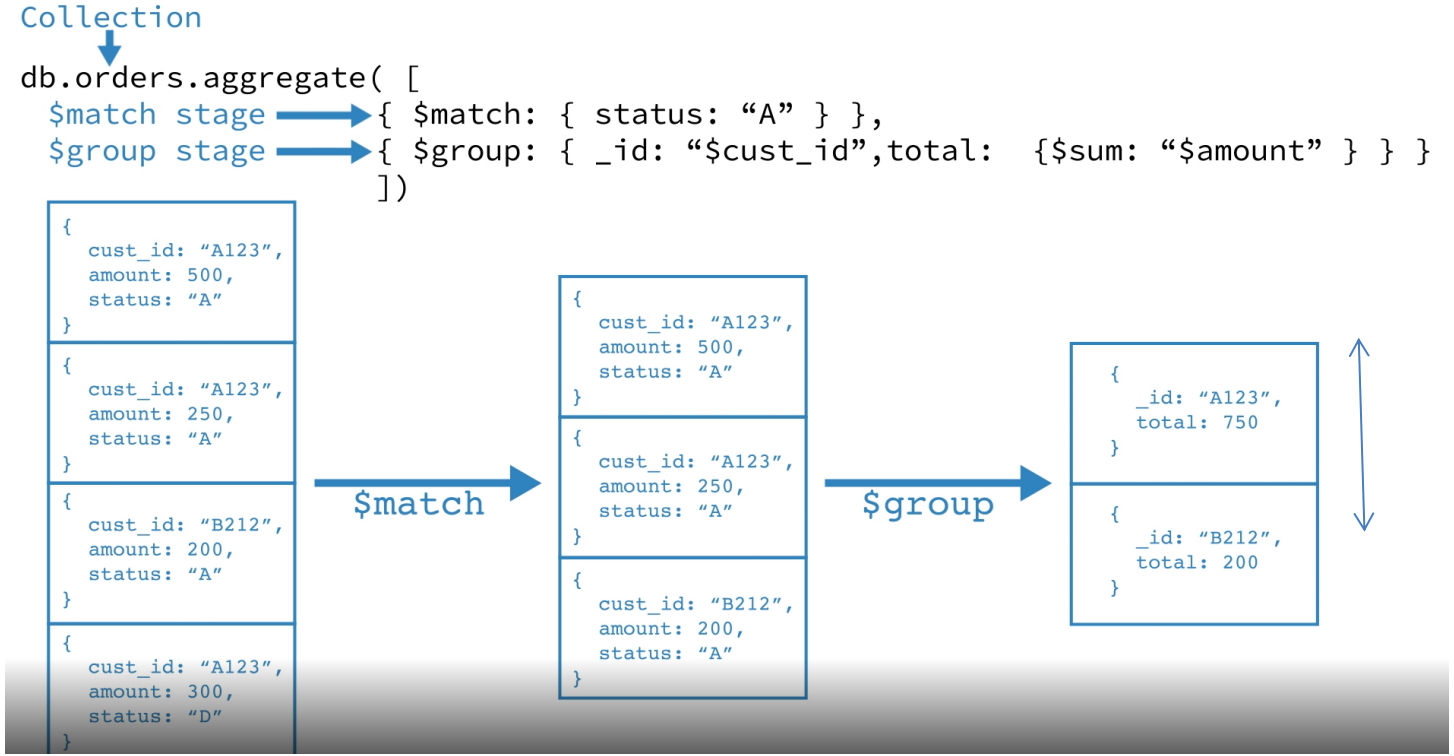
MongoDB Pipeline



MongoDB Pipeline

$sort

, {$sort : {total : 1}}



MongoDB Pipeline

```
db.orders.aggregate([{$match: {status: "A"}}, {$group: {_id: "$cust_id", total: {$sum: "$amount"}}}, {$sort : {total : 1}}])
```

```
/* 1 */
{
    "_id" : "B123",
    "total" : 200
}

/* 2 */
{
    "_id" : "A123",
    "total" : 750
}
```

MongoDB Pipeline

$unwind desagregar arrays, por cada elemento del array crea un nuevo documento duplicando los demás datos.

```
/* 1 */
{
   "_id" : ObjectId("5eaa6b2f804aacfaf16311ce"),
   "nombre" : "roger",
   "deportes" : [ "rally", "basquet”]
}

/* 2 */
{
   "_id" : ObjectId("5eaa6b5d804aacfaf1631bce"),
   "nombre" : "jordi",
   "deportes" : [ "egames” ]
}
```

```
/* 3 */
{
   "_id" : ObjectId("5eaa6b95804aacfaf163286d"),
   "nombre" : "monica",
   "deportes" : ["basquet", "zumba"]
}

/* 4 */
{
   "_id" : ObjectId("5eaa6bb7804aacfaf1632ffd"),
   "nombre" : "jony",
   "deportes" : ["futbol", "egames”]
}
```

MongoDB Pipeline

```
db.getCollection('deportesPreferidos').aggregate([{$unwind : "$deportes"}])
```

deportesPreferidos 0.123 sec.

	_id	nombre	deportes
1	ObjectId("5...	roger	rally
2	ObjectId("5...	roger	basquet
3	ObjectId("5...	jordi	egames
4	ObjectId("5...	monica	basquet
5	ObjectId("5...	monica	zumba
6	ObjectId("5...	jony	futbol
7	ObjectId("5...	jony	egames

MongoDB Pipeline

```
db.getCollection('deportesPreferidos').aggregate([{$unwind : "$deportes"}, {$group: {_id:"$deportes", total: {$sum:1}}}])
```

deportesPreferidos 0.119 sec.

	_id	total
1	rally	1.0
2	basquet	2.0
3	egames	2.0
4	futbol	1.0
5	zumba	1.0

Actualizar: Query Selectors

Name	Description
$eq	Matches values that are equal to a specified value.
$gt	Matches values that are greater than a specified value.
$gte	Matches values that are greater than or equal to a specified value.
$in	Matches any of the values specified in an array.
$lt	Matches values that are less than a specified value.
$lte	Matches values that are less than or equal to a specified value.
$ne	Matches all values that are not equal to a specified value.
$nin	Matches none of the values specified in an array.

Actualizar

- *$set cambia valor*

db.Partidos.update({siglas:{$eq:"PSOE"}},{$set:{programa:"Blabla"}})

- $push: añade un valor especificado a un array.

db.Partidos.update({ _id:1}, { $push:{votos : 89} })

https://docs.mongodb.com/manual/reference/method/db.collection.update/
https://docs.mongodb.com/manual/reference/operator/query/#query-selectors

Buscar

db.collection.find(*<criteria>, <projection>)*

db.Partidos.find({siglas:{$eq:"PSOE"}}, {nombre:1})

Projection = nombre atributo y booleano (1 mostrar y 0 no mostrar, lista blanca o lista negra)

https://docs.mongodb.com/manual/reference/method/db.collection.find/
https://docs.mongodb.com/manual/reference/method/db.collection.find/#find-projection

Buscar – Operadores Lógicos

- $all: Devuelve cierto si el array almacenado como valor del atributo tiene los mismos elementos que el proporcionado en la condición.
- $exists: Para comprobar que el atributo existe en el documento.
- $mod: Para comprobar el resto de una división del valor del atributo por un número.
- $ne: Devuelve cierto si no es igual (non equal)
- $in: Devuelve cierto si el valor está entre alguno de los proporcionados.
- $nin: Contrario de $in.
- $or: Para indicar que se debe cumplir al menos una condición de entre un grupo de condiciones.
- $nor: Contrario de $or.
- $size: Devuelve true si coincide con el número de elementos de un array.
- $type: Para comprobar el tipo del valor almacenado (número, cadena…)

Consulta agregados

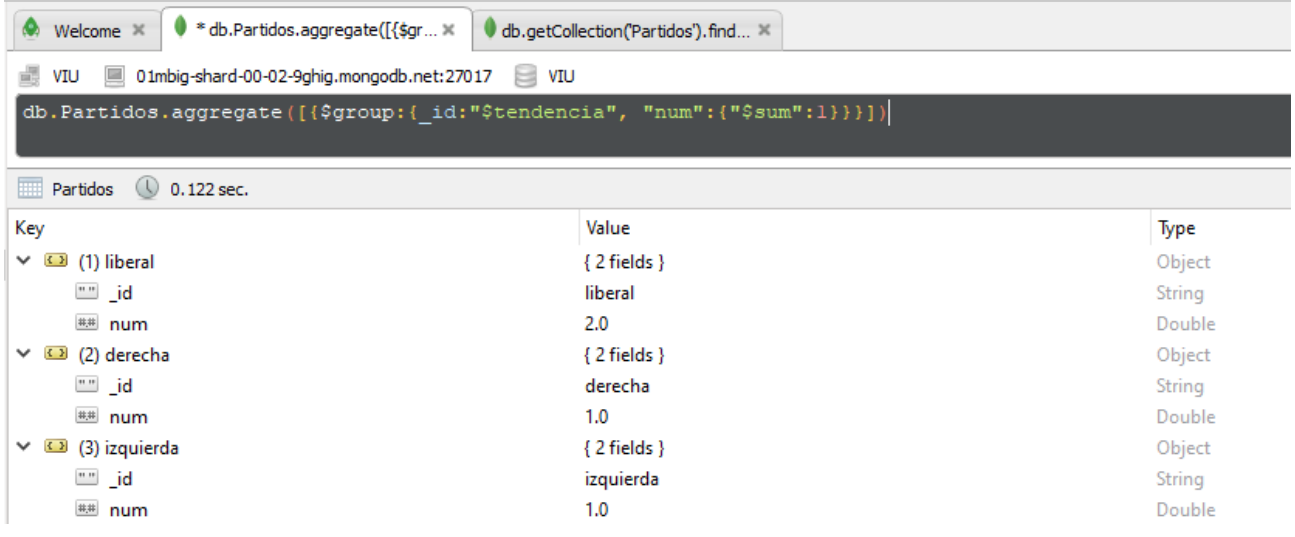
db.collection.aggregate(pipeline)

- $first, $last, $min, $max, $avg, $sum,
- $push o $addtoSet para generar un array con los valores que devuelve, etc.
- $group: indica el atributo por el que se quiere agregar
- $match: criterio a cumplir

Consulta agregados

db.collection.aggregate(pipeline)

db.Partidos.aggregate([{$group:{_id:"$tendencia", "num":{"$sum":1}}}])

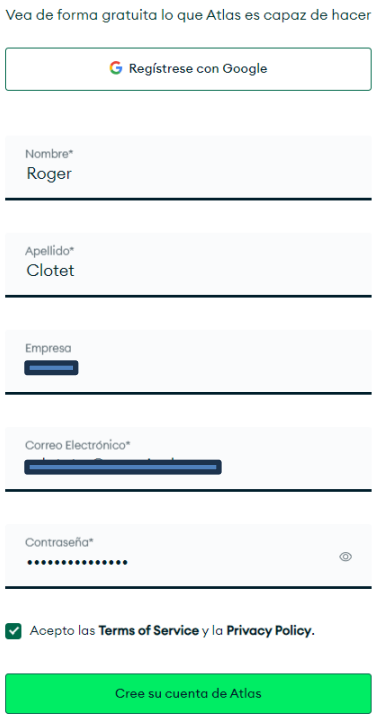


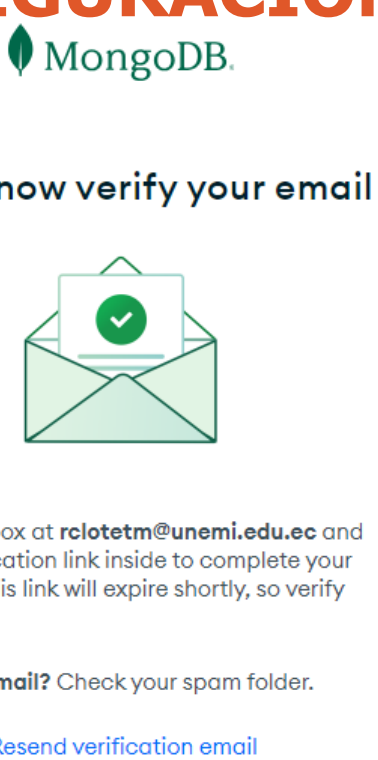
Dudas

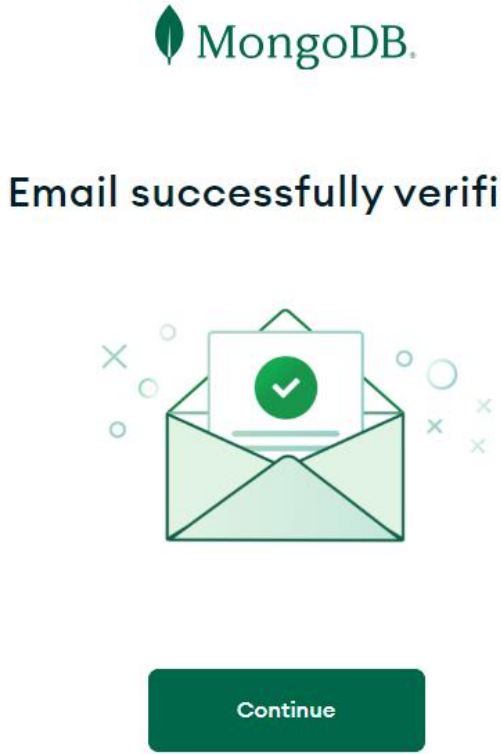
CONFIGURACIÓN

MongoDB Atlas https://www.mongodb.com/

CONFIGURACIÓN







CONFIGURACIÓN

MongoDB.

Enable your multi-factor authentication methods

Secure your account across MongoDB

Secure your MongoDB account across MongoDB Atlas, University, Community, and Support with Multi-Factor-Authentication (MFA). Setting up MFA will help reduce risks from outside threats to your account. Learn More

Setup Backup MFA Methods

In addition, you will also be setting up an additional method in this setup. This is a precaution in the case you lose access to your first MFA enabled device and to avoid being locked out.

Set up now | Remind me later

MongoDB.

Multi-factor authentication setup is now complete

Your MFA methods have now been enabled. You can always add or remove more methods under your Account Application inside of the Security page.

Multi-factor authentication (MFA) methods

In the case that you lose access to your first enabled MFA method, set up another one to avoid being locked out. Learn more

✓	Email	
✓	Okta Verify Mobile App	
○	Security Key/Biometric	SETUP
○	Authenticator App	SETUP
○	SMS	SETUP

Why is this important for me?

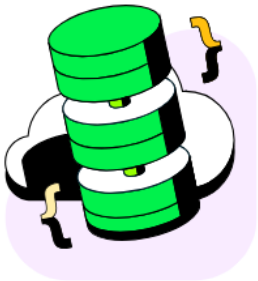
MFA will enhance your account's security by requiring you to identify yourself by more than your email and password.

Learn how this protects your account

Continue

CONFIGURACIÓN

Overview



Create a cluster

Choose your cloud provider, region, and specs.

+ Create

CONFIGURACIÓN

Deploy your cluster

Use a template below or set up advanced configuration options. You can also edit these configuration options once the cluster is created.

M10 $0.09/hour

For production applications with sophisticated workload requirements.

STORAGE	RAM	vCPU
10 GB	2 GB	2 vCPUs

Serverless

For application development and testing, or workloads with variable traffic.

STORAGE	RAM	vCPU
Up to 1 TB	Auto-scale	Auto-scale

M0 Free

For learning and exploring MongoDB in a cloud environment.

STORAGE	RAM	vCPU
512 MB	Shared	Shared

Free forever! Your M0 cluster is ideal for experimenting in a limited sandbox. You can upgrade to a production cluster anytime.

Name
You cannot change the name once the cluster is created.

Cluster0

Automate security setup

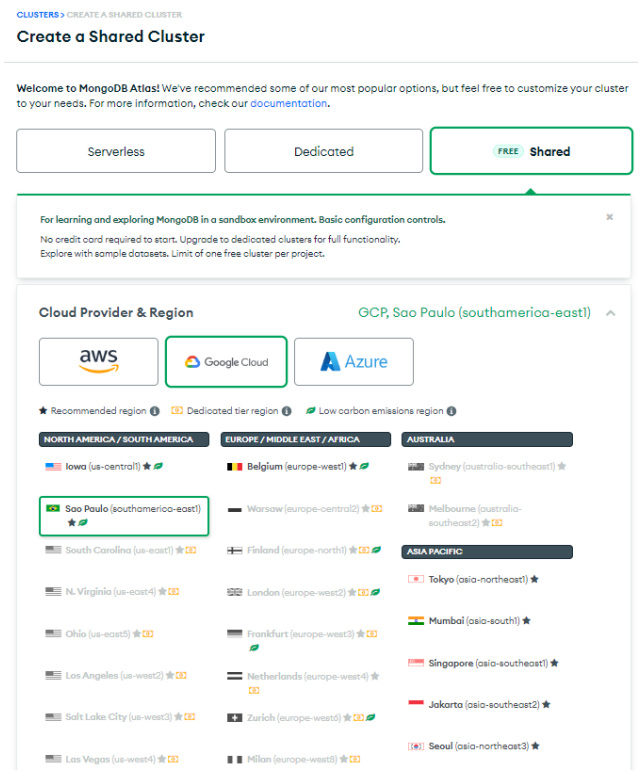
Preload sample dataset

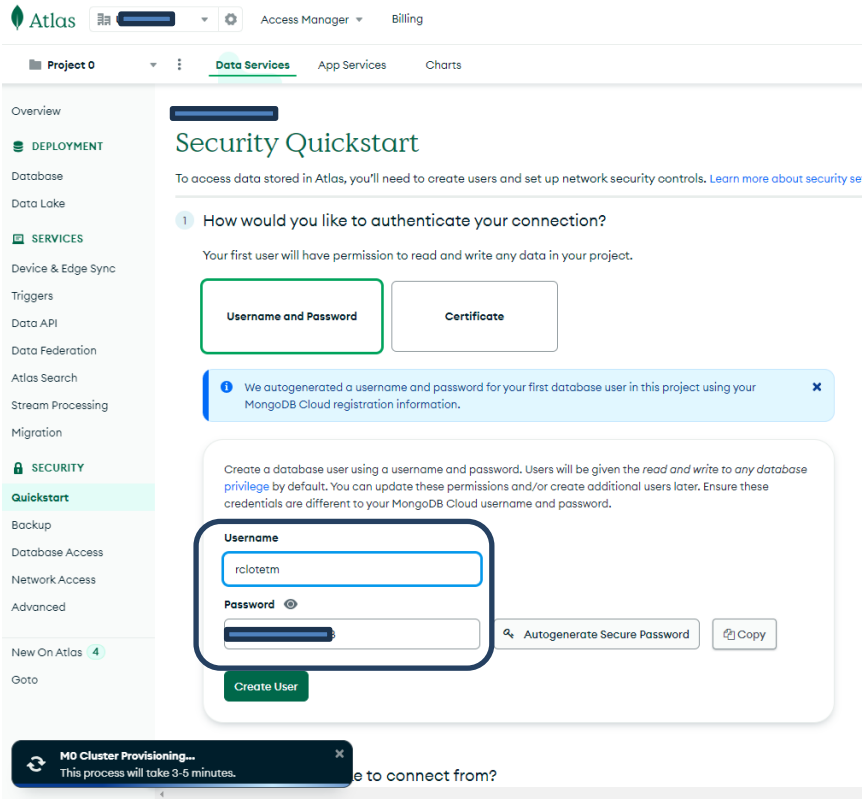
Provider

aws | Google Cloud | Azure

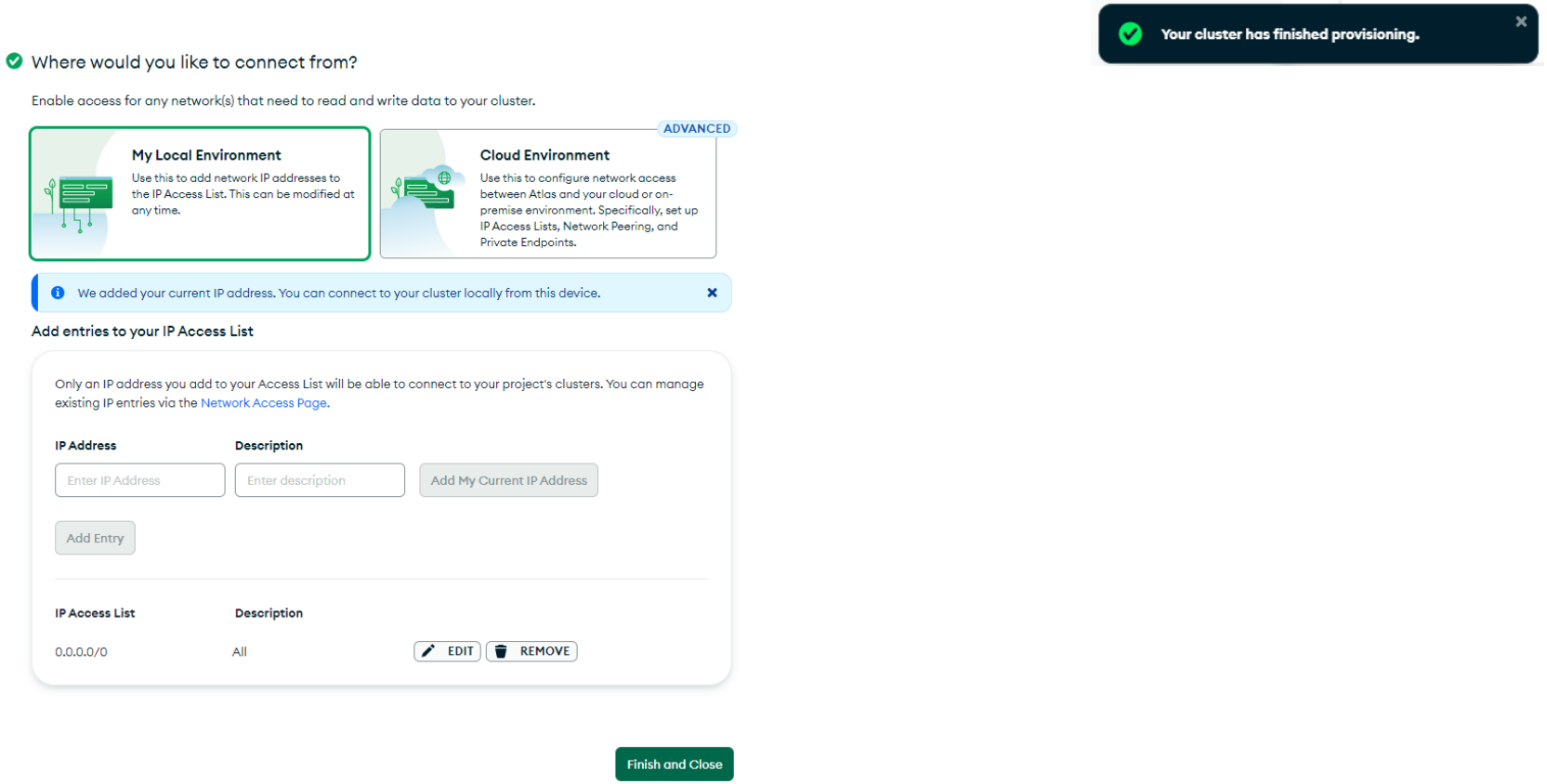
Go to Advanced Configuration | Create Deployment

CONFIGURACIÓN

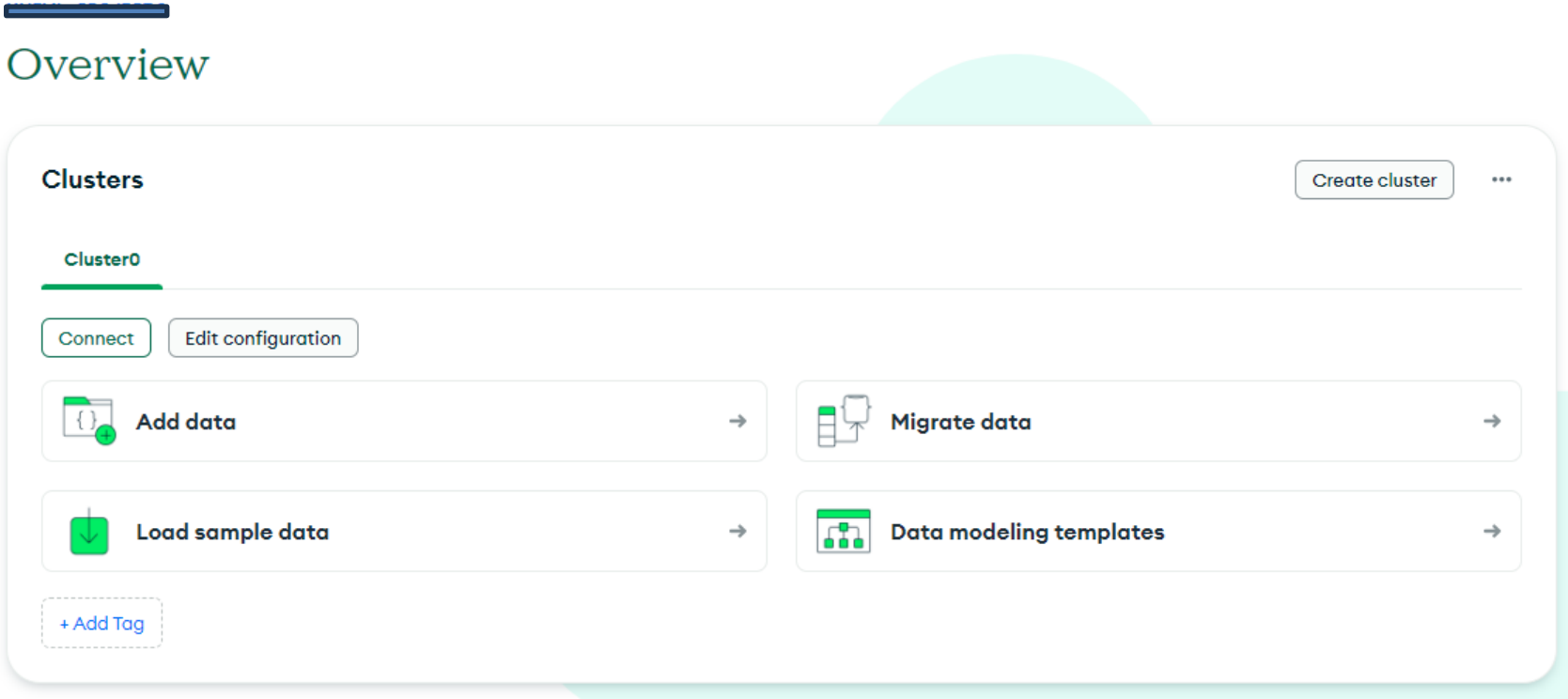




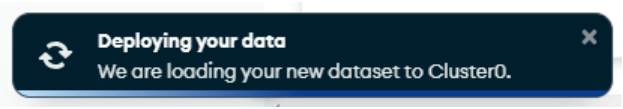
CONFIGURACIÓN

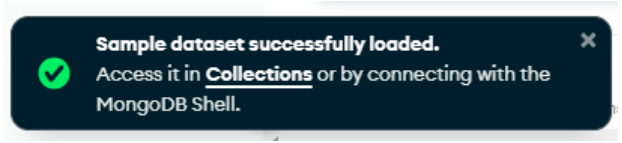


CONFIGURACIÓN



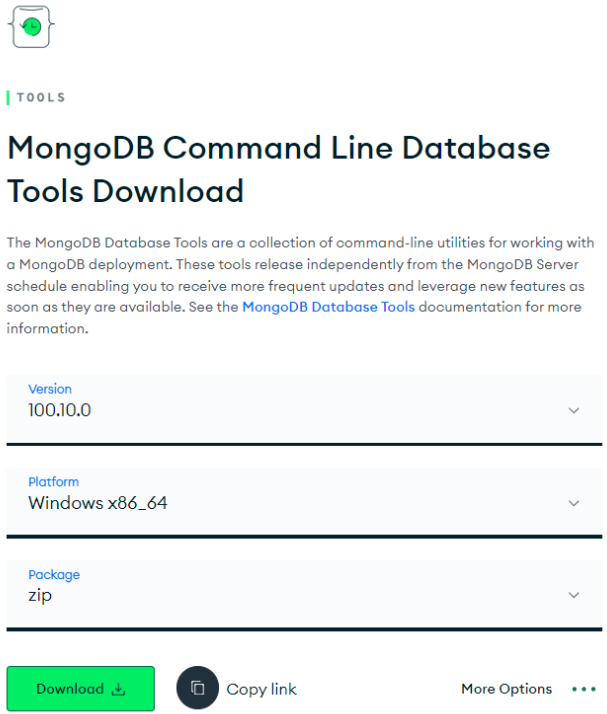
CONFIGURACIÓN





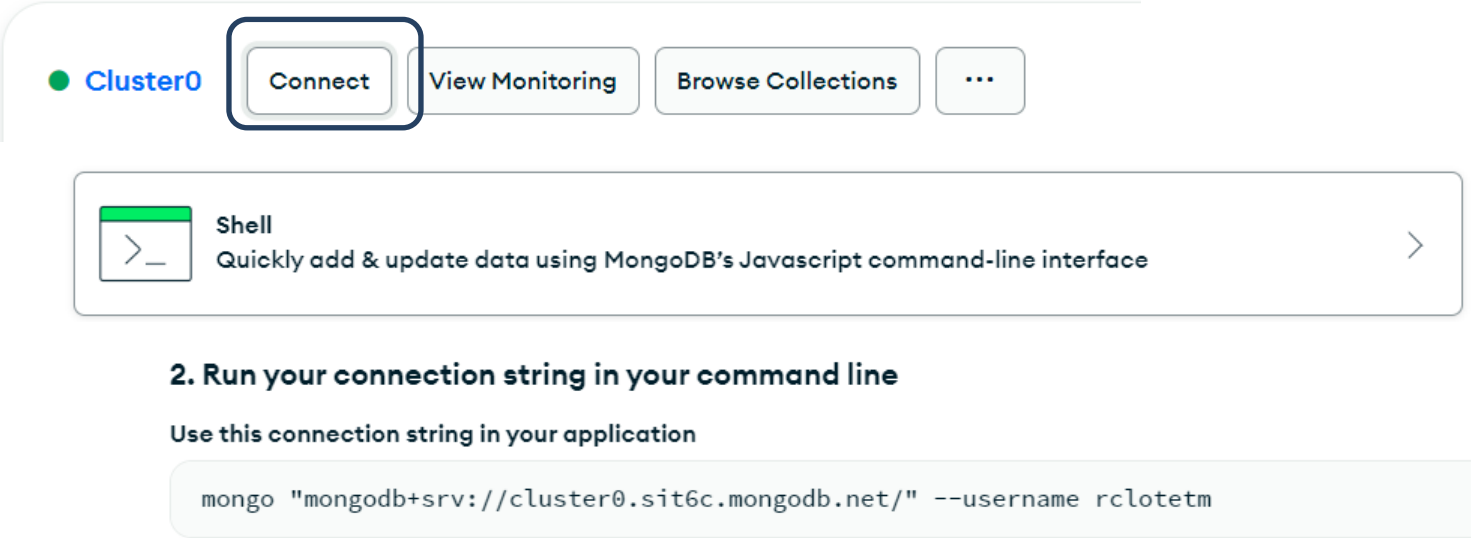
CONFIGURACIÓN

https://www.mongodb.com/try/download/database-tools



CONFIGURACIÓN

Mongoimport



CONFIGURACIÓN

Mongoimport

mongoimport "**mongodb+srv://cluster0.sit6c.mongodb.net/**" --**username rclotetm** --**password ?????** --db VIU --collection **tweets** --file tweets_UNEMI_2.json --type json

mongoimport "mongodb+srv://cluster0.sit6c.mongodb.net/" --username rclotetm --password ????? --db VIU –collection **cuentas** --file cuentas_UNEMI_2.json --type json

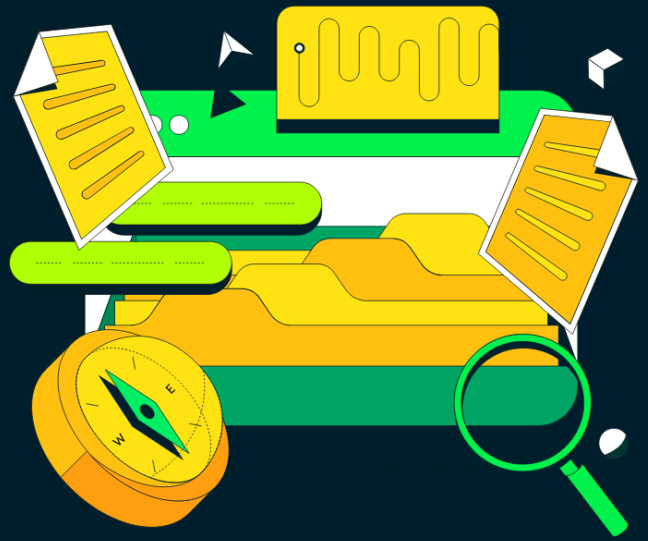
CONFIGURACIÓN

Mongoimport

```
[redacted] mongoimport "mongodb+srv://cluster0.sit6c.mongodb.net/" --username rclotetm --password
[redacted]_2.json --type json
2024-09-22T18:49:06.048-0500    connected to: mongodb+srv://cluster0.sit6c.mongodb.net/
2024-09-22T18:49:09.051-0500    [#.......................] UNEMI.tweets 3.06MB/42.4MB (7.2%)
2024-09-22T18:49:12.050-0500    [###.....................] UNEMI.tweets 6.75MB/42.4MB (15.9%)
2024-09-22T18:49:15.054-0500    [###.....................] UNEMI.tweets 6.75MB/42.4MB (15.9%)
2024-09-22T18:49:18.063-0500    [#######.................] UNEMI.tweets 13.2MB/42.4MB (31.2%)
2024-09-22T18:49:21.060-0500    [#########...............] UNEMI.tweets 16.6MB/42.4MB (39.1%)
2024-09-22T18:49:24.060-0500    [#########...............] UNEMI.tweets 16.6MB/42.4MB (39.1%)
2024-09-22T18:49:27.052-0500    [###########.............] UNEMI.tweets 20.2MB/42.4MB (47.5%)
2024-09-22T18:49:30.063-0500    [##############..........] UNEMI.tweets 23.8MB/42.4MB (56.0%)
2024-09-22T18:49:33.056-0500    [##############..........] UNEMI.tweets 23.8MB/42.4MB (56.0%)
2024-09-22T18:49:36.063-0500    [################........] UNEMI.tweets 26.6MB/42.4MB (62.9%)
2024-09-22T18:49:39.055-0500    [##################......] UNEMI.tweets 30.3MB/42.4MB (71.5%)
2024-09-22T18:49:42.054-0500    [##################......] UNEMI.tweets 30.3MB/42.4MB (71.5%)
2024-09-22T18:49:45.062-0500    [###################.....] UNEMI.tweets 34.2MB/42.4MB (80.5%)
2024-09-22T18:49:48.056-0500    [###################.....] UNEMI.tweets 34.2MB/42.4MB (80.5%)
2024-09-22T18:49:51.051-0500    [#####################...] UNEMI.tweets 38.9MB/42.4MB (91.6%)
2024-09-22T18:49:54.049-0500    [########################] UNEMI.tweets 42.4MB/42.4MB (100.0%)
2024-09-22T18:49:57.053-0500    [########################] UNEMI.tweets 42.4MB/42.4MB (100.0%)
2024-09-22T18:49:58.483-0500    [########################] UNEMI.tweets 42.4MB/42.4MB (100.0%)
2024-09-22T18:49:58.483-0500    11991 document(s) imported successfully. 0 document(s) failed to import.
```

CONFIGURACIÓN

Compass. The GUI for MongoDB.



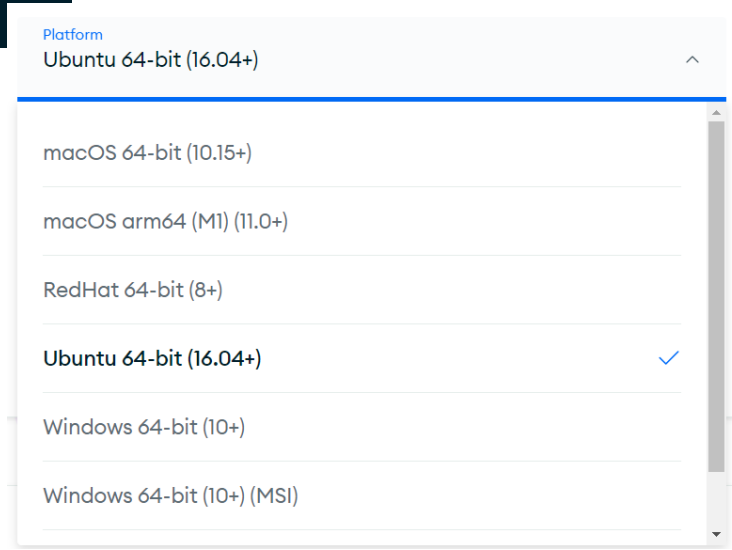
Trabaje fácilmente en con sus datos en Compass, la interfaz gráfica de usuario desarrollado por y para MongoDB. Compass proporciona de todo, desde el análisis de esquemas hasta la optimización de índices y pipelines de agregación en una única interfaz centralizada.

- Compass está disponible gratuitamente para descargar y usar
- Disponible en Linux, Mac o Windows

CONFIGURACIÓN

Compass. The GUI for MongoDB.

https://www.mongodb.com/try/download/compass



CONFIGURACIÓN

Access your data through tools

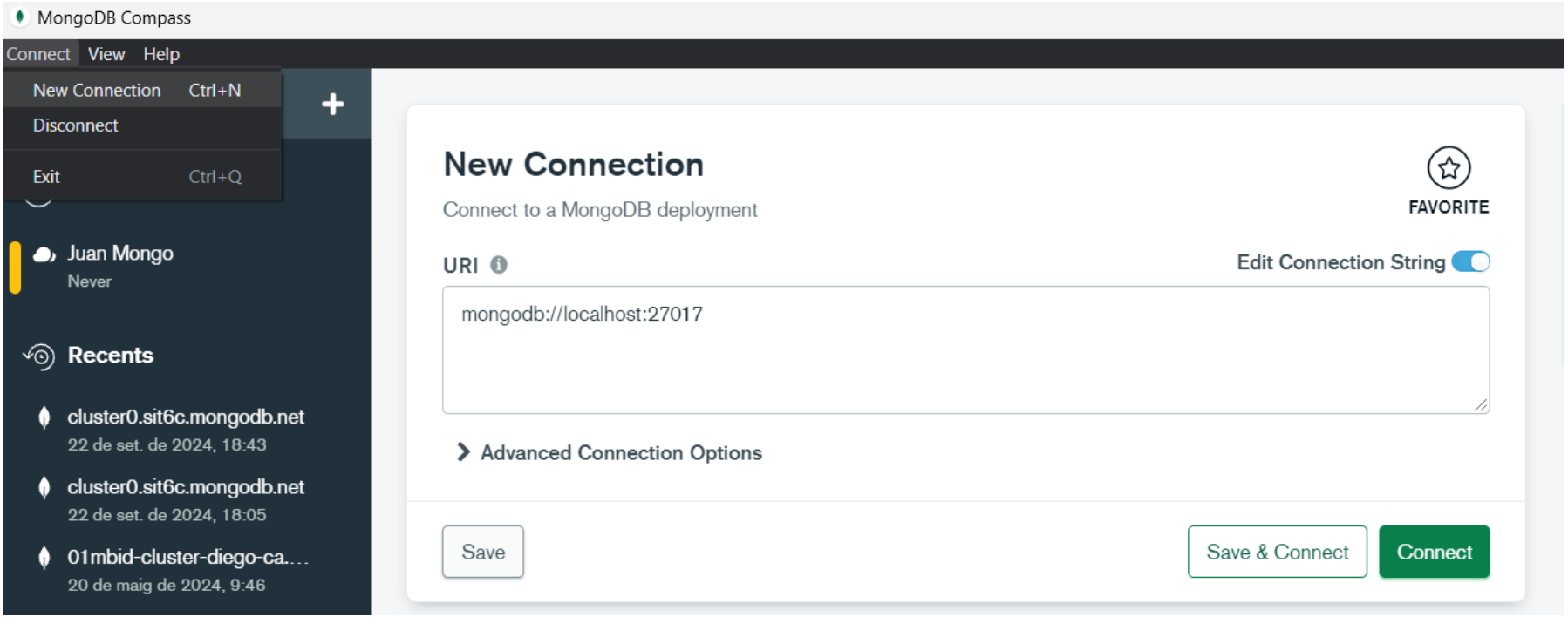
Compass
Explore, modify, and visualize your data with MongoDB's GUI

2. Copy the connection string, then open MongoDB Compass

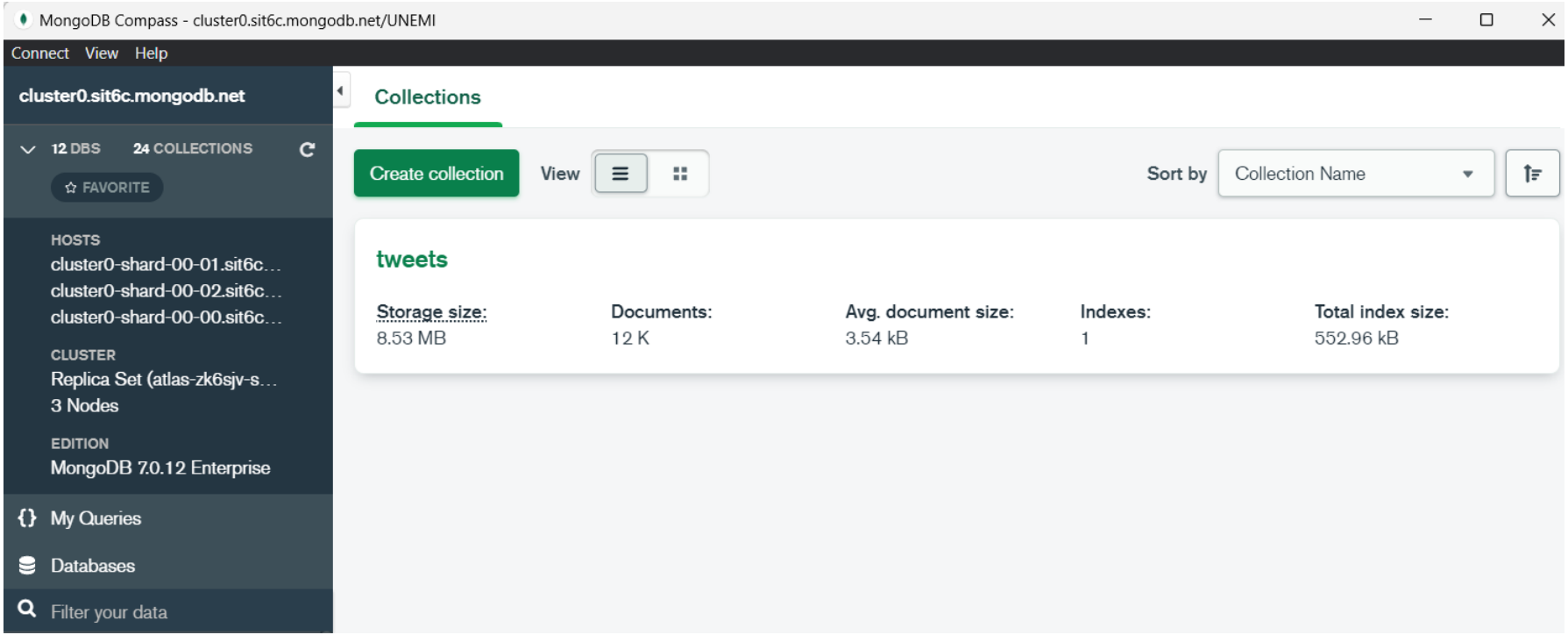
Use this connection string in your application

```
mongodb+srv://rclotetm:<db_password>@cluster0.sit6c.mongodb.net/
```

CONFIGURACIÓN



CONFIGURACIÓN



CONFIGURACIÓN

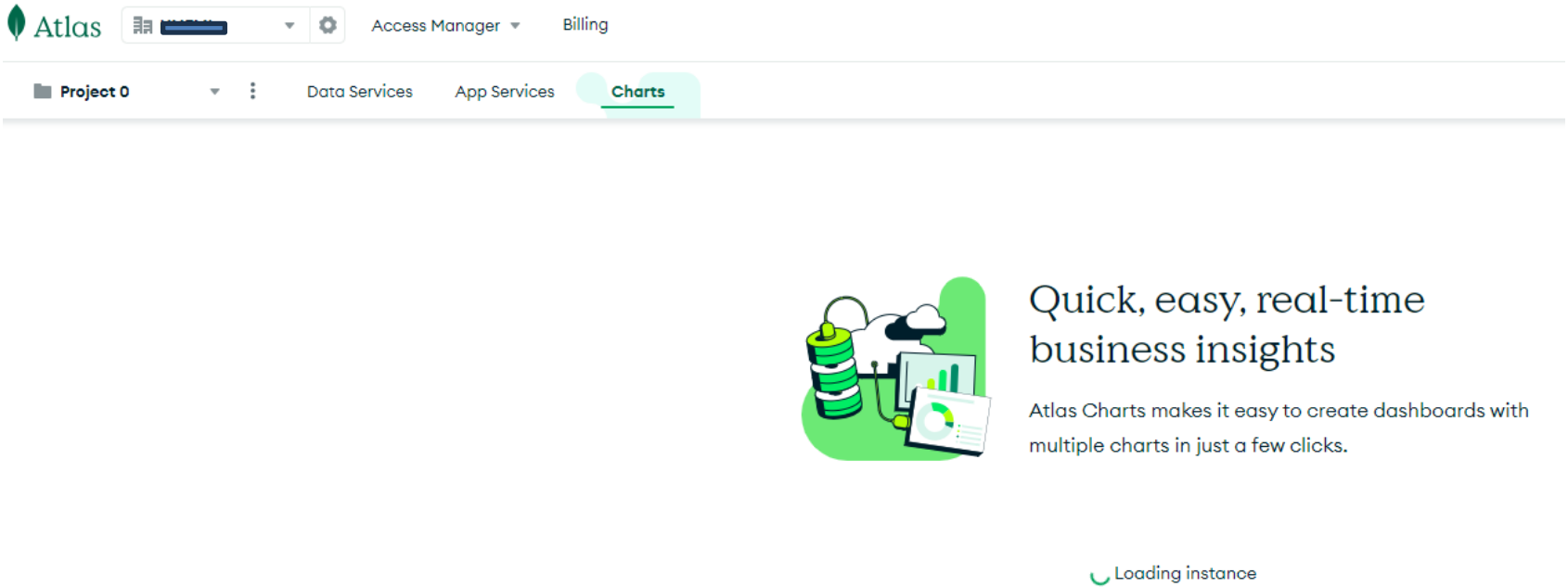
Tweet Object

https://developer.x.com/en/docs/x-api/v1/data-dictionary/object-model/tweet

Visor JSON de un documento (tweet)

https://codebeautify.org/jsonviewer

CONFIGURACIÓN



CONFIGURACIÓN

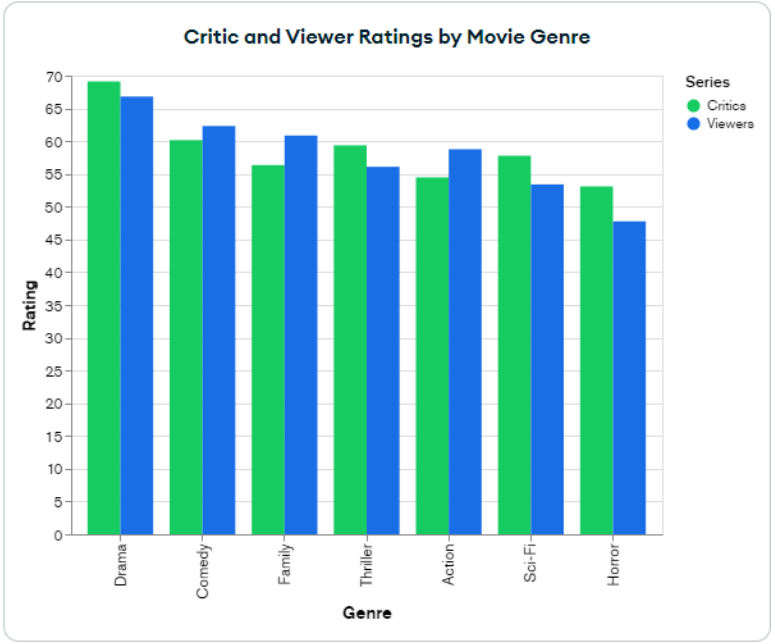
MongoDB.

Hi Roger!
Welcome to Atlas Charts

Charts is the best way to create visualizations of your data.

- Zero setup – Build a chart from any collection in your project
- Create, share and collaborate on charts and dashboards
- Embed charts and dashboards into applications
- Completely free to get started and no user-based pricing

Start



01MBID Fundamentos de la tecnología Big Data

CONFIGURACIÓN

MongoDB.

Where would you like to start?

Chart builder

Quickly create a chart with data in your project or sample data.

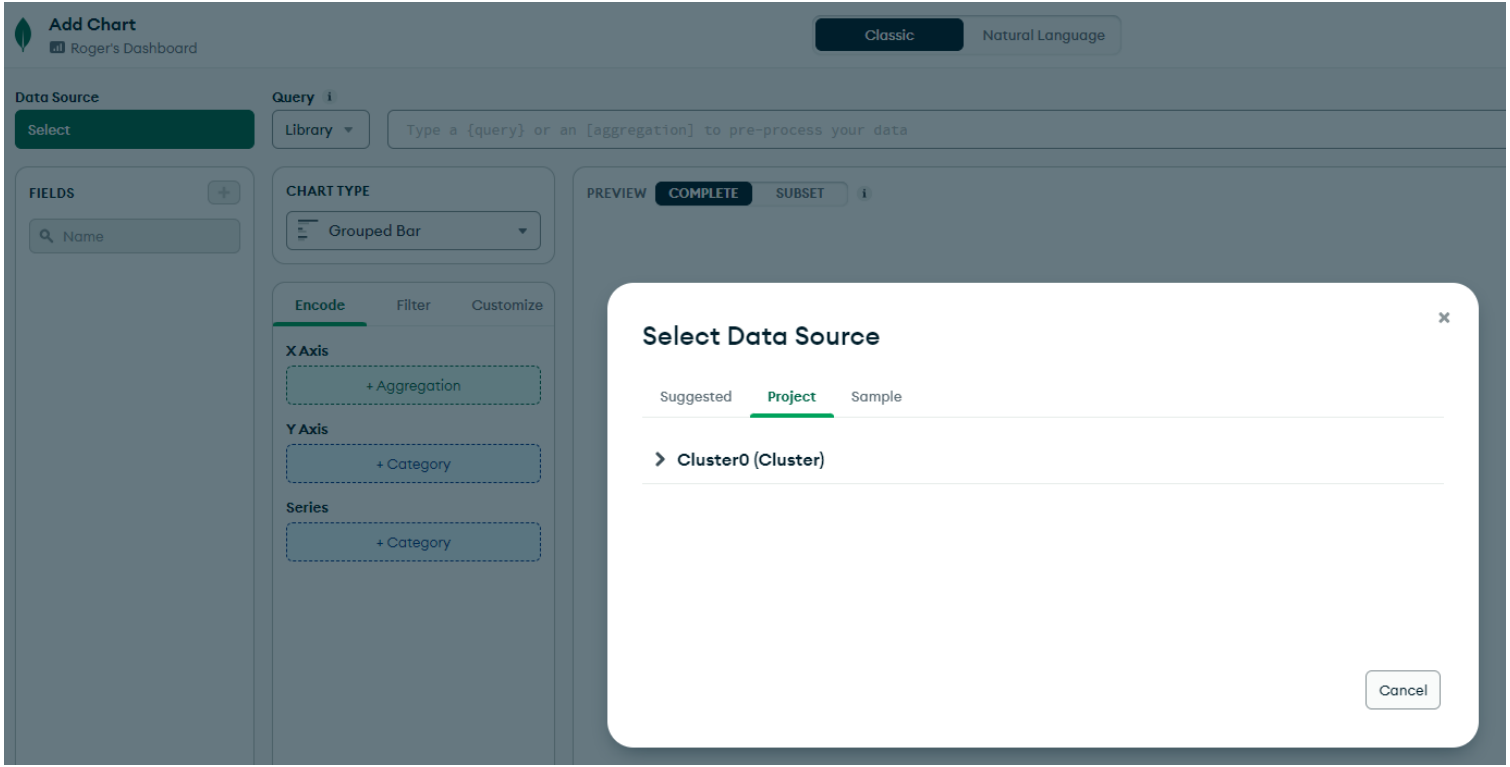
Select

Sample dashboard

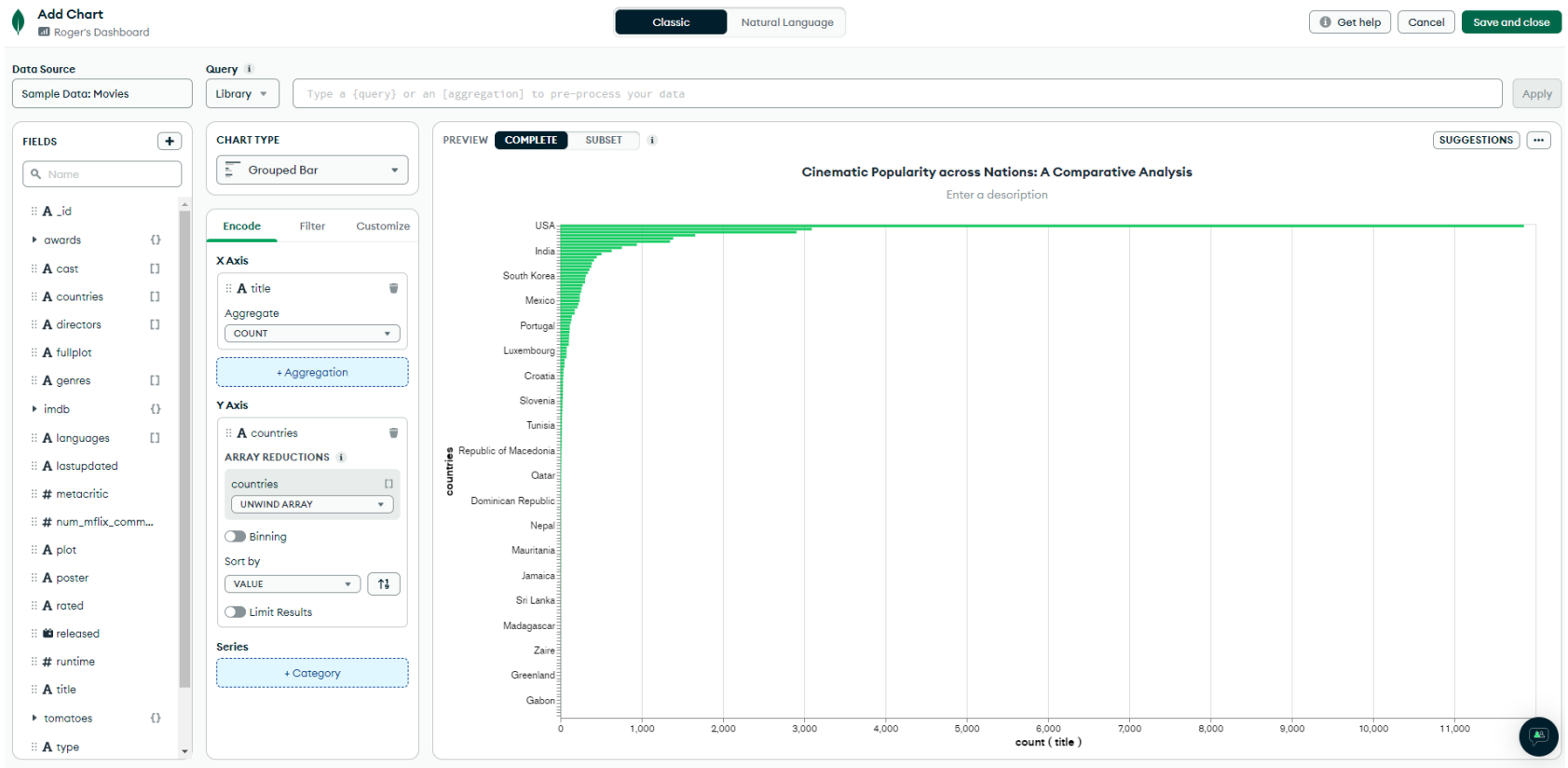
Explore the Atlas Charts features with the sample dashboard.

Select

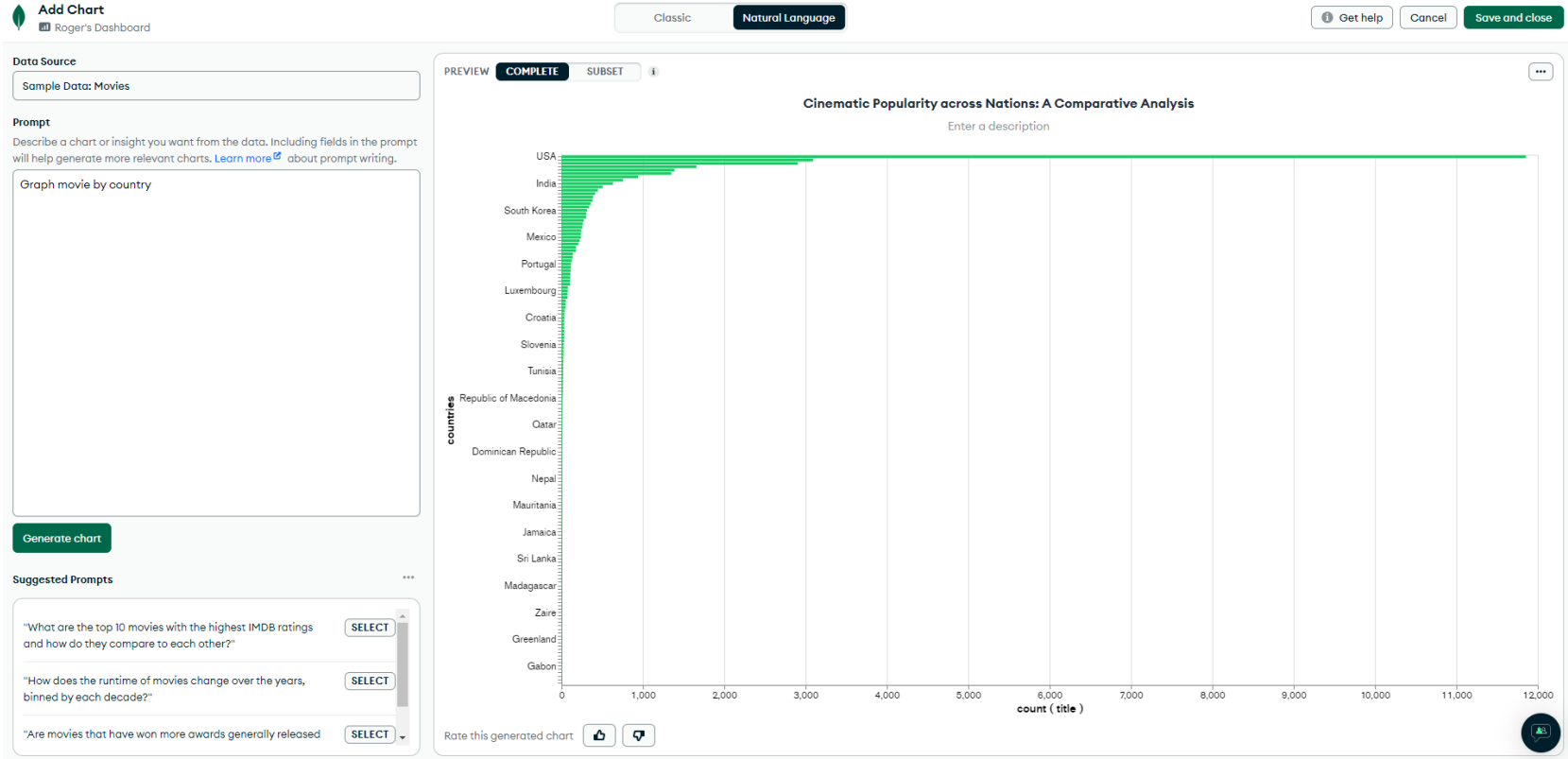
CONFIGURACIÓN



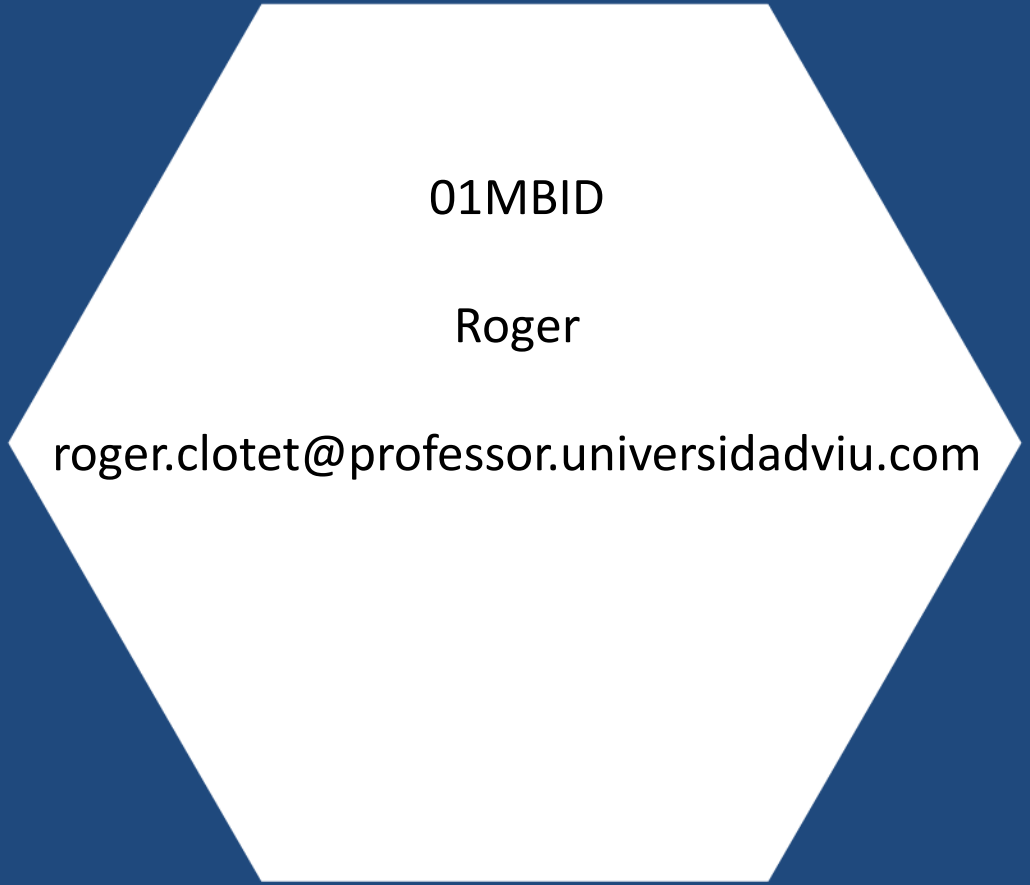
CONFIGURACIÓN



CONFIGURACIÓN



Dudas


01MBID
Roger
roger.clotet@professor.universidadviu.com

Gracias

Universidad Internacional de Valencia