# Índice

[**Índice 1**](#_iyxuaao52t6o)

[**Bases de datos orientadas a documentos 2**](#_spsq54aegac4)

[Explicación del tema 2](#_4xq53oewaa23)

[Datos estructurados 2](#_hmsgvl1dtj1)

[Conformación 2](#_lkr3qlobuo2v)

[Formas de almacenamiento de información 3](#_6at0asd3lbx2)

[Formas de interrelacionar los datos 3](#_qtahkigbenc6)

[Tecnología elegida y sus características 5](#_6mo5qysypr3w)

[**Datasets 6**](#_j4tohkbhz89y)

[Caracterización de la fuente de datos 7](#_ntfi09eo2rwo)

[**Bibliografía 8**](#_56obo5ugbanh)

# 

# Bases de datos orientadas a documentos

# Explicación del tema

El modelo de datos documental está fuertemente orientado a agregados, dado que una base de datos consiste en un conjunto de agregados denominados «documentos». Las bases de datos documentales se caracterizan porque definen un conjunto de estructuras y tipos permitidos que pueden ser almacenados, y además es posible acceder a la estructura del agregado, teniendo como ventaja que se consigue más flexibilidad en el acceso. En este sentido se pueden realizar consultas a la base de datos según los campos del agregado, se pueden recuperar partes del agregado en vez del agregado completo, y además se pueden crear índices basados en el contenido del agregado. Dado que cada agregado tiene asociado un identificador, es posible realizar búsquedas del estilo clave-valor.

Una base de datos orientada a documentos es un tipo de sistema de gestión de bases de datos (DBMS) diseñado para almacenar, recuperar y administrar datos en forma de documentos. En lugar de utilizar tablas y filas como en las bases de datos relacionales, las bases de datos orientadas a documentos almacenan datos en documentos individuales, que pueden ser en formatos como JSON (JavaScript Object Notation), BSON (Binary JSON), XML (Extensible Markup Language) u otros formatos similares.

La naturaleza flexible, semi estructurada y jerárquica de los documentos y las bases de datos de documentos permite que evolucionen según las necesidades de las aplicaciones.

Las bases de datos de documentos ofrecen capacidades de distribución integradas. Puede escalar horizontalmente en varios servidores sin afectar al rendimiento, lo que también es rentable. Además, las bases de datos de documentos proporcionan tolerancia a errores y disponibilidad a través de la replicación integrada.

El modelo de documentos funciona bien con casos de uso como la administración de contenido, los catálogos, la administración de sensores y más. Para cada caso de uso, cada documento es único y evoluciona con el tiempo.

## Datos estructurados

Las bases de datos NoSQL orientadas a documentos permiten el almacenamiento de datos no estructurados o semi-estructurados haciendo uso de clusters de almacenamiento. Un cluster de almacenamiento se refiere a un conjunto de servidores interconectados que trabajan juntos para almacenar y administrar los datos de la base de datos. En este caso, los datos están organizados en documentos, que pueden ser estructuras de datos complejas, como objetos JSON, BSON o XML.

En un cluster de almacenamiento, cada servidor en el conjunto, también conocido como nodo, tiene su propia capacidad de almacenamiento y procesamiento. Los datos se distribuyen entre estos nodos para permitir una mayor escalabilidad y tolerancia a fallos. Esto significa que si un nodo falla, los datos pueden ser recuperados de otros nodos en el cluster.

Los clusters de almacenamiento en bases de datos NoSQL están diseñados para manejar grandes volúmenes de datos de manera eficiente, permitiendo una alta disponibilidad y rendimiento escalable para aplicaciones que requieren acceso a datos rápidos y flexibles.

## Conformación

## Formas de almacenamiento de información

Una base de datos orientada a documentos es un tipo especial de almacén de valores clave en el que las claves solo pueden ser cadenas. Además, el documento está codificado usando estándares como JSON o lenguajes relacionados como XML. También puede almacenar archivos PDF, archivos de imagen o documentos de texto directamente como valores.

Estructura de documentos JSON

La notación de objetos de JavaScript (JSON) es un formato abierto de intercambio de datos legible tanto por los humanos como por las máquinas. Los desarrolladores pueden usar documentos JSON en su código y guardarlos directamente en la base de datos de documentos.

JSON representa los datos de tres maneras:

Valor clave

Los pares clave-valor se registran entre corchetes. La clave es una cadena y el valor puede ser cualquier tipo de datos, como entero, decimal o booleano. Por ejemplo, un valor-clave simple es {"year": 2013}.

Matriz

Una matriz es una colección ordenada de valores definidos entre corchetes izquierdo ([) y derecho (]). Los elementos de la matriz están separados por comas. Por ejemplo, {"fruit": ["apple","mango"]}.

Objetos

Un objeto es un conjunto de pares clave-valor. Básicamente, los documentos JSON permiten a los desarrolladores incrustar objetos y crear pares anidados. Por ejemplo, {"address": {"country": "USA","state": "Texas"}}.

El documento JSON incluso puede tener una matriz con objetos JSON dentro de ella. Por lo tanto, las bases de datos orientadas a documentos le permiten crear una jerarquía de niveles ilimitados de objetos JSON incrustados. Depende totalmente de usted el esquema que quiera darle a su almacén de documentos.

## Formas de interrelacionar los datos

En una base de datos NoSQL orientada a documentos, los datos se relacionan mediante la estructura de documentos, la jerarquía de datos dentro de esos documentos, el uso de referencias o subdocumentos, y la indexación de campos. Esta flexibilidad en el modelado de datos permite adaptarse mejor a los casos de uso y requisitos específicos de las aplicaciones modernas.

Jerarquía de datos: Los documentos pueden contener datos jerárquicos, lo que significa que un documento puede tener otros documentos anidados dentro de él. Esto permite modelar datos complejos de una manera natural y flexible.

Referencias y subdocumentos: En lugar de utilizar relaciones entre tablas como en las bases de datos relacionales, las bases de datos NoSQL orientadas a documentos a menudo utilizan referencias o subdocumentos para representar relaciones entre datos. Por ejemplo, en lugar de almacenar toda la información del cliente en una tabla y referenciarla desde otra tabla, en una base de datos orientada a documentos, podríamos almacenar los datos del cliente directamente dentro de un documento y luego referenciar ese documento desde otras colecciones según sea necesario.

Claves de indexación: Las bases de datos NoSQL orientadas a documentos suelen permitir la indexación de campos dentro de los documentos. Esto facilita la búsqueda y recuperación eficiente de datos basada en los valores de esos campos.

# Tecnología elegida y sus características

Info sobre mongodb : <https://datascientest.com/es/mongodb-todo-sobre-la-base-de-datos-nosql-orientada-a-documentos>

info sobre AWS DocumentDB:

<https://aws.amazon.com/es/documentdb/>

Info sobre Apache CouchDB

[¿Qué es CouchDB? | IBM](https://www.ibm.com/mx-es/topics/couchdb)

### 

# Datasets

Datasets propuestos para elegir:

| Integrante | Sa |
| --- | --- |
| Nombre | Climate Change Tweets Ids |
| Resúmen | Este conjunto de datos contiene los identificadores de tweets de 39.622.026 tweets relacionados con el cambio climático. Se recopilaron entre el 21 de septiembre de 2017 y el 17 de mayo de 2019 desde la API de Twitter mediante Social Feed Manager. Existe una brecha en la recopilación de datos entre el 7 de enero de 2019 y el 17 de abril de 2019.  Los tweets se recopilaron utilizando el método de filtro/estado POST de Twitter Stream API, utilizando el parámetro track con las siguientes palabras clave: #climatechange, #climatechangeisreal, #actonclimate, #globalwarming, #climatechangehoax, #climatedeniers, #climatechangeisfalse, #globalwarminghoax, #climatechangenotreal, climate change, global warming, climate hoax  Debido al tamaño de la colección, la lista de identificadores se divide en archivos de 10 millones de líneas cada uno, con un identificador de tweet en cada línea. Hay un archivo README.txt que contiene documentación adicional sobre cómo se recopilaron los tweets.  El método de búsqueda/estados GET admite la recuperación del tweet completo para una identificación de tweet (conocido como hidratación). Se pueden utilizar herramientas como Twarc o Hydrator para hidratar los tweets.  Según la Política para desarrolladores de Twitter, los identificadores de tweets pueden compartirse públicamente con fines académicos; los tweets pueden no hacerlo. |
| Tamaño |  |
| URL | https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2F5QCCUU&ref=hackernoon.com |

| Integrante | Sa |
| --- | --- |
| Nombre | What A Waste Global Database |
| Resúmen | What a Waste es un proyecto global para agregar datos sobre la gestión de residuos sólidos de todo el mundo. Esta base de datos presenta las estadísticas recopiladas a través del esfuerzo, que cubre casi todos los países y más de 330 ciudades. Las métricas incluidas cubren todos los pasos de la cadena de valor de la gestión de residuos, incluida la generación, composición, recolección y eliminación de residuos, así como información sobre tarifas y financiamiento de los usuarios, el sector informal, estructuras administrativas, comunicación pública e información legal. La información presentada es la mejor disponible basada en un estudio de la literatura actual y conversaciones limitadas con agencias y autoridades de residuos. Si bien puede haber variaciones en las definiciones y la calidad de los informes para puntos de datos individuales, las tendencias generales deberían reflejar la realidad global. Se anotan todas las fuentes y estimaciones. |
| Tamaño |  |
| URL | https://datacatalog.worldbank.org/search/dataset/0039597 |

| Integrante | yar |
| --- | --- |
| Nombre | Monthly Airline Delays by Airport, 2003-2016 |
| Resúmen | Retrasos mensuales de aerolíneas por aeropuerto, 2003-2016.  Incluye aeropuertos, fechas, estadísticos, carriers, vuelos, minutos de demora, etc.  Solo para aeropuertos de EE. UU., no internacionales. |
| Tamaño |  |
| URL | https://think.cs.vt.edu/corgis/datasets/json/airlines/airlines.json |

| Integrante | An2 |
| --- | --- |
| Nombre | Yelp Dataset |
| Resúmen | This dataset is a subset of Yelp's businesses, reviews, and user data. It was originally put together for the Yelp Dataset Challenge which is a chance for students to conduct research or analysis on Yelp's data and share their discoveries. In the most recent dataset you'll find information about businesses across 8 metropolitan areas in the USA and Canada. |
| Tamaño | 4 gb |
| URL | [Yelp Dataset (kaggle.com)](https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset/data) |

| Integrante |  |
| --- | --- |
| Nombre |  |
| Resúmen |  |
| Tamaño |  |
| URL |  |

# Caracterización de la fuente de datos

Diccionario de datos

# Bibliografía

* <https://aws.amazon.com/es/nosql/document/>
* El Gepeto 😎→ este tiene que ser: 🦗