Edición: 2025/2026 **Arquitectura de Datos** **Grado en Ingeniería Informática** **ASUNTO:** Práctica Obligatoria 2.2 – Migración a MongoDB de acuerdo a Casos de Uso	uc3m Universidad Carlos III de Madrid

1. CONTEXTO

La empresa StreamIt S.A. ofrece un servicio de televisión bajo demanda (películas y series). Cada cliente tiene un contrato y recibe facturas mensuales que incluyen consumos de contenidos.

En esta Parte 2 de la práctica se trabajará sobre un dataset de facturas que contiene inconsistencias y datos sucios. El objetivo es extraer, transformar y cargar los datos en MongoDB, así como explotarlos mediante consultas de agregación que den respuesta a distintos casos de uso definidos.

2. DATASET DE PARTIDA

Se proporciona un único fichero (datafiles.zip) con una colección de documentos de facturas, donde cada registro representa la factura mensual de un cliente e incluye:

- Información relativa a tarifas, consumos y contrato.
- Datos del cliente.
- Información embebida sobre los contenidos consumidos (películas, series), junto con estadísticas de uso.

Es fundamental que exploréis detenidamente el dataset antes de trabajar con él.

A partir de esa exploración, deberéis identificar los principales problemas de calidad de los datos, tanto estructurales como de contenido, y dejar constancia de ellos.

Ejemplo de estructura de documento de factura mensual:

```
{
    "_id": "MK23180/11/1/2016",
    "billing": "January 2016",
    "TOTAL": 16.65,
    "Client": { "customer code": "60/17997520/11T", ... },
    "contract": {
          "contract ID": "MK23180/11", ...,
          "product": { "Reference": "Free Rider", ... },
          ...
    },
    "Movies": [ {
          "Date": "15/01/2016", ...,
          "License": { "Date": "15/01/2016" ... },
          "Details": {
                "Year": 2002, ...,
                "Director": { "Name": "Paul W.S. Anderson", ... },
                "Cast": {
                      "Facebook likes": 17902,
```

```
                    "Stars": [ { "Player": "Colin Salmon", ... }, ... ] },
                ...  },
          "Viewing PCT": 99
          }, ...
    ],
    "Series": [ {
          "Date": "09/01/2016", ...,
          "License": { "Date": "09/01/2016", ... },
          "Viewing PCT": 98
          }, ...
    ],
    "charge date": "03/05/17",
    "dump date": "19/12/15"
}
```

3. LIMPIEZA Y NORMALIZACIÓN DE DATOS

El dataset contiene errores e inconsistencias que deben corregirse en una fase de limpieza previa a su explotación.

Cada grupo deberá:

- Explorar los datos para identificar problemas de calidad (fechas en formatos no estándar, campos heterogéneos, valores ausentes, duplicados, etc.).
- Definir y aplicar reglas de consistencia propias, justificando todas las decisiones en la memoria.
- Transformar los datos para ajustarlos al modelo de información objetivo, representado en el diagrama UML proporcionado.

La fase de limpieza debe realizarse exclusivamente en MongoDB mediante operaciones de actualización (update, updateMany) y pipelines de agregación.

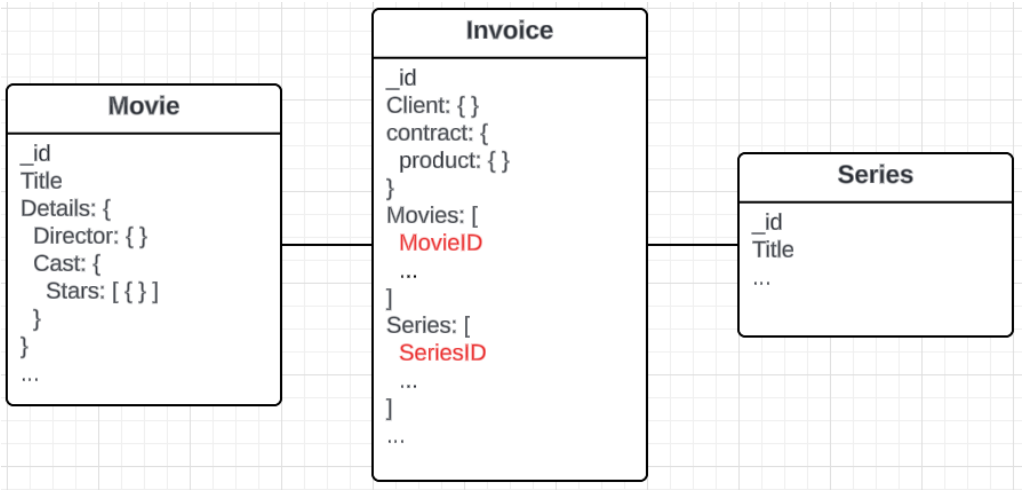
4. REESTRUCTURACIÓN DEL MODELO DE DATOS

Por motivos de diseño del clúster distribuido y alineado con el modelo de negocio (ofrecer métricas de consumo a clientes y segmentar la venta de publicidad en plataformas), a partir de la colección única de facturas se debe generar:

- Una colección de películas, con la información relativa a los contenidos de tipo película.
- Una colección de series y temporadas, con la información correspondiente.

La colección de facturas debe quedar simplificada, sin contener información redundante de películas y series, siendo sustituida por las referencias necesarias a los documentos de las nuevas colecciones.

Se proporciona un diagrama UML de referencia con el modelo de información esperado.



El tipado de los atributos deberá ser el siguiente:

- Fechas y horas: *ISODate* (todos los campos que representen momentos temporales deben convertirse desde string a formato estándar de MongoDB).
- Importes monetarios: *Decimal128* (para facturas, totales, precios u otros valores económicos).
- Códigos e identificadores: *String* (por ejemplo: códigos de cliente, contrato, factura), salvo que se usen como referencias a documentos de otras colecciones, en cuyo caso es *ObjectId*.
- Nombres, títulos y descripciones: *String* (se recomienda homogeneizar mayúsculas y tildes).
- Listas de valores: *array<String>* (ej.: apellidos, géneros, palabras clave).
- Porcentajes o métricas de consumo: *Int32* (0–100) o *Double* si se requiere precisión decimal.
- Duraciones, contadores y cantidades enteras: *Int32* (ej.: duración en minutos, número de temporadas, episodios, etc.).

Este tipado de atributos deberá cumplirse tras la limpieza y transformación, y servirá de referencia para los esquemas de validación correspondientes.

5. VALIDACIÓN DE ESQUEMAS

Definir e implementar los esquemas de validación para cada colección creada, aplicando:

- Las reglas de consistencia proporcionadas.
- Los criterios adicionales definidos por el grupo.

6. CONSULTAS DE AGREGACIÓN

A partir del nuevo diseño de la base de datos, implementar los siguientes casos de uso en MongoDB:

Q1) Serial Lovers: detectar clientes que han visionado al menos una temporada completa de una serie.

Q2) Apellidos comunes: obtener el apellido más frecuente por país de contrato.

Q3) Actores populares: listado con el top-5 de actores más vistos en España.

Q4) Clásicos modernos: top-10 anual de películas del siglo XX más vistas.

Q5) Facturación mensual: total de ingresos en un mes.

Q6) Consumo mensual: total mensual de visionados (series vs películas), incluyendo duración en horas.

Q7) Películas exitosas: número total de visionados y duración acumulada para una película concreta (añadiendo campos).

Q8) Fracaso en series: promedio de porcentaje visto por usuario en cada serie para analizar abandono.

7. DOCUMENTACIÓN A ENTREGAR Y FECHA

Se entregará un **fichero comprimido en formato ZIP**, que deberá incluir los siguientes ficheros:

1. Memoria (único documento en PDF, con fuente Arial 11 e interlineado sencillo, nombrado **PO22_numGrupoReducido_numGrupoPracticas.pdf**) que contenga los siguientes apartados:
   - Limpieza y normalización de datos. La extensión será de hasta 3 páginas.
   - Reestructuración del modelo de datos. La extensión será de hasta 3 páginas.
   - Validación de esquemas. La extensión será de hasta 3 páginas.
   - Consultas de agregación. La extensión será de hasta 6 páginas.

   Se deberán incluir las capturas de pantalla de la salida de las operaciones en MongoDB en los apartados correspondientes.

   La declaración de uso de IA se incluirá como anexo a la memoria, y la extensión será a demanda.

   El código se incluirá de forma separada, por lo que en la memoria solo se deben incluir explicaciones y justificaciones junto a las capturas de pantalla.
2. Código (4 ficheros en formato TXT, con el código utilizado en cada uno de los 4 primeros apartados anteriores, incluyendo con los comentarios descriptivos pertinentes). Los ficheros se nombrarán de la siguiente forma:
   - **PO22_numGrupoReducido_numGrupoPracticas_1_limpieza.txt**
   - **PO22_numGrupoReducido_numGrupoPracticas_2_reestructuracion.txt**
   - **PO22_numGrupoReducido_numGrupoPracticas_3_esquemas.txt**
   - **PO22_numGrupoReducido_numGrupoPracticas_4_agregaciones.txt**

En la tarea habilitada en Aula Global se entregará un fichero comprimido en formato ZIP (con nombre **PO22_numGrupoReducido_numGrupoPracticas.zip**), el cual deberá incluir los 5 ficheros anteriores, antes del lunes 10 de noviembre a las 09:00 horas por un solo miembro del equipo como representante.

8. RÚBRICA DE EVALUACIÓN

Para que la práctica sea evaluada, es imprescindible que se ajuste al formato y a la extensión fijada, además de incluir la declaración de uso de IA. En caso contrario, la calificación será de 0 puntos.

Criterio	Sobresaliente (100%)	Notable (75%)	Suficiente (50%)	Insuficiente (25%)	Deficiente (0%)	Peso
Fase de limpieza y normalización de datos	Se aplican correctamente todas las reglas y se documenta cada decisión	Mayoría de reglas aplicadas, con leves inconsistencias	Varias reglas omitidas o mal aplicadas	Se aplica solo parcialmente, sin justificar	No se aplica limpieza	20%
Reestructuración del modelo de datos	Las nuevas colecciones se generan con la estructura indicada, y la colección de partida queda simplificada solo con las referencias	Las nuevas colecciones se generan con la estructura indicada, y la colección de partida mantiene redundancias	Las nuevas colecciones se generan aunque con estructura diferente, y la colección de partida mantiene redundancias	Las nuevas colecciones se generan incorrectamente, y la colección de partida mantiene redundancias	No se realiza reestructuración	20%
Validación de esquemas	Se implementan completamente los esquemas para todas las colecciones	Se implementan todos los esquemas pero no de forma completa	Se implementan solo algunos esquemas y de forma incompleta	Se implementa solo un esquema y es incompleto	No se realiza	10%
Consultas de agregación	Consultas correctas y bien justificadas	Correctas con fallos menores	Incompletas o incorrectas parcialmente	Se han intentado, pero sin resultado	No se incluyen	40%
Presentación del documento	Clara, estructurada, sin errores, con capturas adecuadas	Clara y estructurada, con leves errores	Poco clara, con errores frecuentes	Desordenada o difícil de seguir	Desordenada, ilegible	10%