Edición: 2025/2026 **Arquitectura de Datos** **Grado en Ingeniería Informática** **ASUNTO:** Práctica Obligatoria 3 – Cassandra	uc3m Universidad Carlos III de Madrid

1. INTRODUCCIÓN

El objetivo de esta práctica es diseñar e implementar un modelo de datos en Apache Cassandra que permita resolver de forma óptima un conjunto de consultas analíticas sobre un dataset basado en los pasajeros del Titanic.

2. DATASET

Se trabajará con un dataset de Titanic, compuesto por dos ficheros CSV.

El primero recoge información personal de los pasajeros, incluyendo su identificador único, edad, nombre, sexo y variables familiares como el número de padres, hijos o hermanos a bordo.

El segundo archivo contiene datos relacionados con el billete y el trayecto: clase en la que viajaba el pasajero, precio pagado, número de cabina, puerto de embarque y estado de supervivencia.

Aunque los datos son relativamente sencillos, presentan variabilidad, valores ausentes y distintos niveles de granularidad. Por ello, es recomendable realizar una exploración previa antes del diseño del modelo. Podéis trabajar con los datos originales, pero cualquier decisión de limpieza (eliminación de filas, imputación de edades, etc.) debe documentarse.

3. DISEÑO

El modelo de Cassandra debe realizarse teniendo en cuenta las siguientes restricciones de diseño:

- No se permite `ALLOW FILTERING`.
- No se permiten índices secundarios.

Por tanto, las consultas del apartado 5 deben resolverse exclusivamente mediante el modelo diseñado. Para ello, se deben definir las tablas que consideréis necesarias para dar soporte eficiente a dichas consultas y justificar detalladamente la elección de las claves de partición y claves de clustering.

En la justificación debéis relacionar las decisiones de diseño con los riesgos evitados: particiones demasiado grandes, cardinalidades problemáticas, desequilibrios en el anillo, etc.

> **NOTA.** Podéis añadir nuevos atributos en las tablas de Cassandra (por ejemplo, campos derivados o rangos de edad), siempre que estén justificados y faciliten la resolución eficiente de las consultas.

El preprocesado fuera de Cassandra se utilizará principalmente para integrar y limpiar los datos de los ficheros originales, así como para generar, en su caso, columnas derivadas. **No obstante, no se acepta que las consultas se resuelvan únicamente mediante ficheros ya agregados o precalculados donde Cassandra actúe solo como un contenedor de resultados finales (p. ej., tablas que solo requieran un SELECT * sin un diseño real de claves de partición y clustering).**

4. IMPLEMENTACIÓN DISEÑO Y CARGA

4.1. Implementación

De acuerdo con el diseño realizado a casandra, hacer la implementación de las tablas en CQL.

4.2. Carga de datos

El proceso de carga debe garantizar que los datos de ambos ficheros iniciales pueden utilizarse para resolver todas las consultas del apartado 5.

a) Preparación externa a Cassandra

Debido a que Cassandra no soporta operaciones de tipo JOIN, es necesario integrar externamente los datos de los dos ficheros. Esta integración puede realizarse con Python/Pandas, Spark, MongoDB u otras tecnologías equivalentes.

El resultado será uno o varios ficheros CSV para poblar las tablas diseñadas en el apartado 4.1.

b) Carga mediante `COPY FROM`

Cada tabla se poblará importando su CSV correspondiente mediante el comando `COPY ... FROM`.

c) Verificación

Tras la carga, se deberá validar el número total de registros importados en cada tabla, comprobar la ausencia de claves nulas y revisar, al menos de forma aproximada, el equilibrio de las particiones.

5. CONSULTAS

Resolved las siguientes consultas. Adjuntad capturas de pantalla con las salidas. Si lo consideráis necesario, podéis incluir reflexiones sobre problemas encontrados o alternativas de diseño consideradas.

Se valorarán evidencias básicas de rendimiento (tiempos aproximados, número de filas por partición entre distintas soluciones o de la solución final, etc.).

5.1. Supervivientes por clase.

Pasajeros supervivientes filtrados por clase (1, 2 o 3). La consulta debe ejecutarse sin full scan.

5.2. Pasajeros por puerto ordenados por edad.

Pasajeros cuyo puerto de embarque sea S, C o Q, ordenados de menor a mayor edad.

5.3. Mujeres supervivientes por clase.

Obtener información que permita analizar cuántas mujeres supervivientes hay en cada clase.

5.4. Pasajeros por rango de edad.

Pasajeros cuya edad se encuentre entre un rango dado, ordenados por edad.

5.5. Análisis por puerto y supervivencia.

Para cada puerto de embarque, obtener información que permita comparar el volumen de pasajeros y la distribución de supervivientes y no supervivientes.

5.6. Análisis por clase, edad y supervivencia.

Obtener, dado un rango de edad, devolver para cada clase el número de pasajeros supervivientes y no supervivientes dentro de ese rango. El resultado debe permitir comparar fácilmente la supervivencia entre clases.

6. CUESTIONES

Responded a las siguientes cuestiones:

6.1. ¿Cómo habrían cambiado las consultas si hubierais permitido ALLOW FILTERING? ¿serían eficientes?

6.2. ¿Por qué los índices secundarios pueden parecer útiles, pero no son adecuados para entornos de producción en Cassandra?

6.3. ¿Qué impacto tiene un mal diseño de claves sobre aspectos como el rendimiento, la compactación, la distribución del anillo o las lecturas excesivas?

7. ENTREGABLES

Se entregará un archivo comprimido en formato ZIP que deberá incluir lo siguiente:

7.1. Memoria

Único documento en PDF (fuente Arial 11, interlineado sencillo), nombrado como PO3_numGrupoReducido_numGrupoPracticas.pdf. La memoria deberá contener:

- Apartado 3 (máximo 3 páginas).
- Apartado 4 (código del 4.1 y del 4.2 (apartados b y c)) (máximo 4 páginas).
- Apartado 5 (máximo 3 páginas).
- Apartado 6 (máximo 2 páginas).

Las capturas y la declaración de uso de IA se incluirán como **anexo**, sin límite de extensión.

7.2. Código del preprocesado y ficheros CSV

Debe incluirse el código del preprocesado realizado fuera de Cassandra, junto con todos los ficheros .csv generados. El archivo deberá nombrarse PO3_numGrupoReducido_numGrupoPracticas_preprocesado.txt y los CSV según el nombre de cada tabla: nombreTabla.csv

7.3. Script completo ejecutable por el profesor

Archivo txt que incluya:

- Creación de las tablas,
- Importación mediante COPY,
- Consultas del número de registros de cada tabla,
- Consultas del apartado 5.

El archivo se nombrará PO3_numGrupoReducido_numGrupoPracticas_consultas.txt

8. FECHA DE ENTREGA

En la tarea habilitada en Aula Global se entregará el fichero comprimido PO3_numGrupoReducido_numGrupoPracticas.zip, antes del viernes 5 de diciembre a las 09:00 horas por un solo miembro del equipo como representante.

9. 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
1. Diseño del modelo (tablas + claves)	Diseño totalmente correcto. Claves de partición y clustering bien elegidas y justificadas con criterios de volumen, acceso y distribución. Se evita hot partitions. Se plantean tablas relacionando con consultas apartado 5	Diseño casi correcto, con pequeñas dudas o supuestos no justificados. Pocas mejoras pendientes.	Hay varios fallos de diseño (p. ej., particiones demasiado grandes o claves mal elegidas), pero la estructura es funcional.	Diseño incorrecto en aspectos críticos: claves mal definidas, uso de claves inadecuadas, diseño incompatible con consultas.	No se presenta un diseño válido o no funciona.	40%
2. Carga del dataset	Carga completa, reproducible y correctamente documentada. Se verifica número total de registros + coherencia básica de campos	Carga correcta, con validación parcial o documentación incompleta.	La carga funciona, pero con errores leves.	La carga es incompleta, inconsistente o no se puede reproducir.	No se realiza la carga o está totalmente mal.	10%
3. Consultas apartado 5	Todas las consultas funcionan, están bien diseñadas	Todas las consultas funcionan, pero alguna presenta estructuras mejorables o justificaciones incompletas.	Las consultas funcionan parcialmente o requieren ALLOW FILTERING	Varias consultas no funcionan o muestran diseños incorrectos	No se incluyen las consultas o no ejecutan.	35%
4. Cuestiones apartado 6	Explicaciones claras, profundas y bien estructuradas. Se demuestra comprensión de: particiones, clustering, compaction, y diseño por query. Se analizan alternativas y por qué se descartan.	Buenas justificaciones, aunque con menor profundidad o algunos vacíos.	Justificaciones breves o superficiales, pero correctas.	Explicaciones confusas o incorrectas.	No se incluyen reflexiones.	15%