Distribución de datos y *Map Reduce*
PEC2

Ejercicio 1 (25%)

Healthy Soils es una empresa multinacional especializada en la optimización de explotaciones agrarias a través de Internet of Things (IoT). Entre las múltiples soluciones que ofrece a sus clientes destaca un dispositivo, alimentado por una batería propia que, montado sobre una placa Arduino Feather M0 equipada con una tarjeta de comunicaciones LoRaWAN (un protocolo de transmisión de datos asíncrono), que incluye los siguientes sensores:

- GS3: este sensor, que se introduce bajo tierra, ofrece una medida de la temperatura, otra de la conductividad eléctrica, y otra de la permeabilidad.
- SHT31: este sensor obtiene la humedad y la temperatura del aire.
- TSL2591: este sensor ofrece tres medidas independientes sobre el grado de luminosidad: la infrarroja, la del espectro entero (full-spectrum) y de la luz visible por el ojo humano.

Se debe tener en cuenta que en el futuro es posible que se incorporen nuevos tipos de sensores a la placa, incrementando el número y variedad de los datos generados. Se desea que esos datos también sean incorporados a la misma base de datos.

Healthy Soils se ha propuesto implantar este dispositivo en miles de explotaciones por todo el mundo. Cada placa enviará los datos de sus sensores conjuntamente, en intervalos de 10 minutos.

Además, para la explotación de los datos almacenados, la empresa creará una API que será consumida por una aplicación web que muestre datos y gráficas de la evolución temporal de los parámetros medidos por los sensores. Esta API también será expuesta, bajo autenticación y autorización, a cada uno de los clientes contratados que deseen utilizarla para visualizar y explotar sus propios datos, por lo que se espera que el número de consultas sea enorme y que éstas se produzcan de forma concurrente desde cualquier parte del mundo.

Se pide:
Estudiar el caso propuesto y responder a las siguientes preguntas:

1. Razonar cuál sería el modelo de almacenamiento de datos más adecuado, mencionando sus ventajas sobre el resto de modelos tratados en el curso.
2. Explicar qué estrategia de fragmentación sería la óptima.
3. Explicar cuál sería la mejor estrategia de replicación de datos.

4. El modelo transaccional más adecuado para los requisitos de la aplicación descrita.

Exponed la solución de forma argumentada en una página y media como máximo.

Los criterios de evaluación para evaluar este ejercicio son los siguientes:

Pregunta	**No logrado (D/C-)**	**Mínimamente logrado (C+)**	**Logrado (B)**	**Logrado de forma sobresaliente (A)**
Pregunta 1	La respuesta es incorrecta o no está justificada.	La respuesta es correcta y está justificada, pero sólo analiza el modelo de datos ganador.	La respuesta es correcta, analiza todos los modelos de datos y está justificada. La justificación en algún caso es incompleta (obvia algunos aspectos relevantes) o discutible.	La respuesta es correcta e indica, por cada tipo de base de datos, su adecuación y su justificación. Los argumentos planteados son completos y adecuados.
Pregunta 2	La respuesta es incorrecta o no está justificada.	La respuesta es correcta y está bien justificada en la mayoría de las características del sistema (disponibilidad, escalabilidad, distribución de carga y localidad de datos).	La respuesta es correcta y está bien justificada en casi todas las características del sistema (disponibilidad, escalabilidad, distribución de carga y localidad de datos).	La respuesta es correcta y está bien justificada para todas las características del sistema (disponibilidad, escalabilidad, distribución de carga y localidad de datos).
Pregunta 3	La respuesta es incorrecta o no está justificada.	La respuesta es correcta y está mínimamente justificada.	La respuesta es correcta, está bien justificada y aborda aspectos geográficos y de localidad de datos.	La respuesta es correcta, está justificada y aborda aspectos geográficos, de localidad de datos y propone una arquitectura de gestión de réplicas adecuada.

Pregunta 4	La respuesta es incorrecta o no está justificada.	La respuesta es correcta y está mínimamente justificada.	La respuesta es correcta, considera los distintos modelos disponibles y está bien justificada.	La respuesta es correcta, considera los distintos modelos disponibles, está bien justificada y justifica de acuerdo al enunciado la decisión tomada (consistencia vs disponibilidad).

Ejercicio 2 (25%)

A partir de la lectura de los apuntes del curso exponed, para cada una de las afirmaciones, si creéis que es cierta o falsa indicando una breve argumentación que justifique vuestra respuesta.

Las afirmaciones en las que no se indique si es cierta o falsa, o bien carezcan de argumentación, se considerarán no válidas. Se **valorará la concisión** (una página y media para las 5 afirmaciones como máximo) y el **uso de referencias** (sección/página del libro de referencia o apuntes del curso) para justificar las respuestas.

Afirmación 1

La fragmentación en una base de datos relacional implica que un conjunto de datos sólo esté disponible en un fragmento y, por lo tanto, el nodo que almacene ese fragmento será el único responsable de sus datos.

Afirmación 2

La principal ventaja del modelo de replicación maestro-esclavo es la consistencia. Nunca es posible que diferentes clientes lean diferentes nodos secundarios y los datos sean diferentes.

Afirmación 3

En el modelo MapReduce, todas las funciones de reducción son combinables.

Afirmación 4

Las bases de datos orientadas a documentos sólo admiten replicación, resultando imposible las técnicas de distribución como sharding.

Afirmación 5

El teorema CAP ha permitido un cambio de paradigma en la programación de la persistencia de información, ya que implica una relajación de las propiedades de transacciones ACID.

Los criterios de evaluación para evaluar este ejercicio son los siguientes:

No logrado (D/C-)	Mínimamente logrado (C+)	Logrado (B)	Logrado de forma sobresaliente (A)
La respuesta es incorrecta o no está justificada/referenciada.	La respuesta es incorrecta, pero la justificación es correcta, demuestra conocimiento sobre el tema y está referenciada.	La respuesta es correcta y está convenientemente justificada/referenciada.	La respuesta es correcta, está correctamente justificada/referenciada y es concisa (ocupa menos de una página y media).

Ejercicio 3 (30%)

Considerar los datos de los atletas que participan en carreras de media y larga distancia en España.

Por un lado, se dispone de información personal de cada atleta, identificado de forma única por el ID del chip con el que participan en las competiciones.

chip_id	nombre	apellido1	apellido2	residente_en
JGK309	Alejandro	García	Rodríguez	Madrid
QNR874	María	Fernández	Pérez	Barcelona
BDM752	Pablo	González	Sánchez	Valencia
TLF931	Laura	López	Martínez	Sevilla
WXP684	Sergio	Torres	Gómez	Bilbao
HVF460	Ana	Ruiz	Jiménez	Barcelona
YTK639	Luis	Moreno	Álvarez	Murcia
EBN206	Carmen	Castro	Romero	Barcelona
IJM125	Javier	Morales	Gutiérrez	Malaga
UAO572	Marta	Navarro	Vázquez	Cordoba

Por otro lado, se cuenta con el registro de las inscripciones de las diferentes carreras oficiales de media/larga distancia en España.

chip_id	tipo_carrera	ciudad_carrera	fecha	tiempo
JGK309	Maratón	Madrid	12-05-2020	02h:43m:18s
QNR874	10Km	Barcelona	23-09-2021	01h:15m:59s
BDM752	5Km	Valencia	04-07-2021	00h:28m:27s
TLF931	Maratón	Sevilla	19-02-2021	03h:26m:45s
WXP684	10Km	Bilbao	29-11-2022	01h:02m:13s
HVF460	Maratón	Zaragoza	08-08-2022	03h:59m:38s
YTK639	Maraton	Murcia	27-01-2020	03h:21m:49s
EBN206	Maraton	Alicante	05-10-2022	03h:48m:02s
IJM125	5Km	Malaga	17-06-2021	00h:22m:56s
ABC123	Maratón	Cordoba	01-04-2021	02h:37m:04s

Se desea identificar el atleta de los residentes en Barcelona con el mejor tiempo de maratón en 2022. El output tendrá que indicar el nombre y apellidos del atleta y el mejor tiempo.

Se aconseja usar dos pasadas map-reduce. Para ello se pide explicar en cada pasada lo que ocurre en cada fase (map, shuffle, reduce).
Concretamente para cada fase:

- Mostrar todos los datos de entrada (input)
- Explicar qué acciones se producen
- Mostrar el resultado producido (output) por esas acciones en todos los datos de entrada.

Es imprescindible explicar claramente la lógica de las acciones realizadas, enseñando en detalle el input y el output de cada fase. Se valorarán la **lógica y la eficiencia del algoritmo** y las explicaciones, y no (pseudo)código.

Los criterios de evaluación para evaluar este ejercicio son los siguientes:

No logrado (D/C-)	Mínimamente logrado (C+)	Logrado (B)	Logrado de forma sobresaliente (A)

La respuesta es incorrecta o no está justificada.	La respuesta es parcialmente incorrecta, pero la justificación es correcta y demuestra conocimiento sobre el tema.	La respuesta es correcta y los diferentes pasos realizados se describen convenientemente y se ejemplifican mediante la evolución de los datos. En algún punto aislado, la respuesta puede ser incompleta (la ausencia de justificación o de cálculo de datos en algún punto) o incorrecta (contener algún error menor).	La respuesta es correcta, presenta y describe los pasos realizados y los ejemplifica correctamente mediante todos los cálculos de los datos desde el principio hasta el resultado final.

Ejercicio 4 (20%)

Considerar los siguientes sistemas:

1. Un sistema de sensores de la calidad del aire de una ciudad (consistencia final en el tiempo, Escrituras más frecuentes que lecturas)
2. Un periódico online que es visitado por lectores de diferentes lugares de un país (consistencia final en el tiempo, Lecturas más frecuentes que escrituras)
3. Un sistema de control aéreo (consistencia fuerte, Escrituras y lecturas de la misma importancia)
4. Un sistema de banca online (consistencia fuerte, Lecturas más frecuentes que escrituras)

Se pide argumentar qué configuraciones de los valores W, R y N encajan mejor con cada caso con respecto al tipo de consistencia (fuerte/final en el tiempo) y al tipo de lecturas y escrituras que son necesarias realizar (Lecturas más frecuentes que escrituras, Escrituras más frecuentes que lecturas o Escrituras y lecturas de la misma importancia).

C1: N=3, W=1, R=1
C2: N=3, W=4, R=2
C3: N=5, W=1, R=3
C4: N=5, W=3, R=1
C5: N=7, W=4, R=5
C6: N=7, W=5, R=4
C7: N=5, W=5, R=1

Para resolverlo:

- Primero identificar y razonar tipo de consistencia, y tipo de lecturas y escrituras.
- A continuación, analizar cuál de las configuraciones dadas encaja con el sistema propuesto.

Los criterios de evaluación para evaluar este ejercicio son los siguientes:

No logrado (D/C-)	**Mínimamente logrado (C+)**	**Logrado (B)**	**Logrado de forma sobresaliente (A)**
La respuesta es incorrecta o no está justificada.	La respuesta es incorrecta en algún aspecto, pero la justificación es correcta, demuestra conocimiento sobre el tema y está referenciada.	La respuesta es correcta y está convenientemente justificada, pero sólo justifica la configuración ganadora o bien el análisis de alguna de las configuraciones es incorrecta.	La respuesta es correcta, está correctamente justificada y aborda todas las configuraciones posibles, indicando para cada una de ellas, su adecuación al problema planteado y el porqué.

Criterios de valoración

Los apartados 1 y 2 tienen un peso del 25% cada uno, y los apartados 3 y 4 tienen un peso del 30% y el 20% respectivamente. Se valorará, para cada apartado, la validez de la solución y la claridad de la argumentación. Cualquier solución no justificada se considerará incompleta.

Formato y fecha de entrega

Tenéis que enviar la PEC al buzón de Entrega y registro de EC disponible en el aula (apartado Evaluación). El formato del archivo que contiene vuestra solución puede ser .pdf,. odt, .doc y .docx. Para otras opciones, por favor, contactar previamente con vuestro profesor colaborador. El nombre del fichero debe contener el código de la asignatura, vuestro apellido y vuestro nombre, así como el número de actividad (PEC2). Por ejemplo nombreapellido1_nosql_pec2.docx. La fecha límite para entregar la PEC2 es el **3 de mayo.**

Propiedad intelectual

Al presentar una práctica o PEC que haga uso de recursos ajenos, se tiene que presentar junto con ella un documento en que se detallen todos ellos, especificando el nombre de cada recurso, su autor, el lugar donde se obtuvo y su estatus legal: si la obra está protegida por el copyright o se acoge a alguna otra licencia de uso (Creative Commons, licencia GNU, GPL etc.). El estudiante tendrá que asegurarse que la licencia que sea no impide específicamente su uso en el marco de la práctica o PEC. En caso de no encontrar la información correspondiente tendrá que asumir que la obra está protegida por el copyright.

Será necesario, además, adjuntar los ficheros originales cuando las obras utilizadas sean digitales, y su código fuente, si así corresponde.