# Distribución de datos y *Map Reduce*
PEC2

## Ejercicio 1 (25%)

En una organización se almacenan y gestionan series temporales (datos indexados en orden temporal) del tipo `{timestamp, métrica, valor, origen, tags}`. Se generan miles de datos por minuto, al final del día hay millones de registros. El tipo de datos generado, es decir los atributos generados y su formato es uniforme y no se esperan cambios a corto plazo. Se supone que los datos son consultados de forma concurrente por multitud de usuarios. Los datos no requieren un almacenamiento indefinido, ya que se utilizan para estudiar el estado de los sistemas.

**Se pide:**
Estudiar la conveniencia de un modelo relacional o uno NoSQL y razonar qué modelo sería el más adecuado.

A partir de la arquitectura de distribución escogida:

1. Indicar las ventajas e inconvenientes que tiene la estrategia escogida respecto a otras por lo que respecta a la fragmentación (partición) de los datos.
2. Indicar las ventajas e inconvenientes que tiene la estrategia escogida respecto a otras por lo que respecta a la replicación de los datos.
3. El modelo transaccional más adecuado para los requisitos de la aplicación descrita.

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

## Ejercicio 2 (25%)

A partir de la lectura del artículo 'Consistency Models of NoSQL Databases' y de los apuntes indica si te parecen ciertas o falsas las siguientes afirmaciones.

Para cada una de las afirmaciones indica si es cierta o falsa, justificando la respuesta mediante lo que has leído en el artículo. En cada justificación deberá **indicar el párrafo del artículo** en la que se sustenta tu argumentación.

**No serán válidas las respuestas que no se justifiquen.**

Se valorará la concisión (una página y media para las 5 afirmaciones como máximo).

### Afirmación 1

Según el teorema CAP se puede afirmar que si una base de datos es CA implica que los datos son consistentes entre todos los nodos (mientras los nodos estén en línea) y que se puede leer/escribir de forma consistente en cualquier nodo puesto que los datos serán los mismos.

### Afirmación 2

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

### Afirmación 3

La consistencia final en el tiempo establece que todas las réplicas llegarán a ser gradualmente consistentes si no hay actualizaciones.

### Afirmación 4

La consistencia fuerte garantiza que una operación de lectura para un dato obtendrá el valor de la última escritura, aunque alguna de estas réplicas pueda tener valores inconsistentes.

## Ejercicio 3 (30%)

Ejecuta el archivo `IntroMapReduce.ipynb` que se entrega junto al enunciado de la PEC siguiendo las instrucciones descritas en el anexo I. A continuación, resuelve los dos supuestos que se proponen. No es necesario que codifiques en Python la solución, basta que expliques tu solución utilizando pseudocódigo, explicaciones textuales así como los datos proporcionados para los resultados.

3.1 Suponer los datos de una red social que consisten en un conjunto de pares del tipo (personaA, personaB) que representan una relación “sigue a” (‘following’) de forma que la personaA sigue a la personaB. Dado el siguiente conjunto de datos,describe el algoritmo MapReduce que calcula el número de seguidores que tiene cada persona.

```
red_social = [('Alicia', 'Benito'), ('Benito', 'Alicia'),
              ('Carlos', 'Benito'), ('Benito', 'Carlos'),
              ('Daniela', 'Enrique'), ('Enrique', 'Francisco'),
              ('Francisco', 'Enrique'), ('Daniela', 'Benito')]
```

3.2 La relación “sigue a” no es simétrica, ya que una persona no tiene porqué seguir a sus seguidores. No obstante, a veces ocurre que una persona sigue a la persona que

lo sigue. Es decir, si PersonaA sigue a PersonaB, entonces PersonaB sigue a PersonaA. Con este ejercicio queremos identificar los pares (PersonaA, PersonaB) que no tienen una relación (PersonaB, PersonaA) definida. Con los mismos datos anteriores, describe el algoritmo MapReduce que permite obtener la lista de las relaciones que cumplen dicha condición.

# Ejercicio 4 (20%)

Suponiendo que debemos configurar el sistema de replicación de una base de datos para que cumpla la condición de consistencia fuerte y se nos plantean distintas alternativas. Sabemos que se reciben muchas solicitudes de lectura y relativamente pocas de escritura. ¿Cuál de las siguientes alternativas de quórum elegirías y por qué?

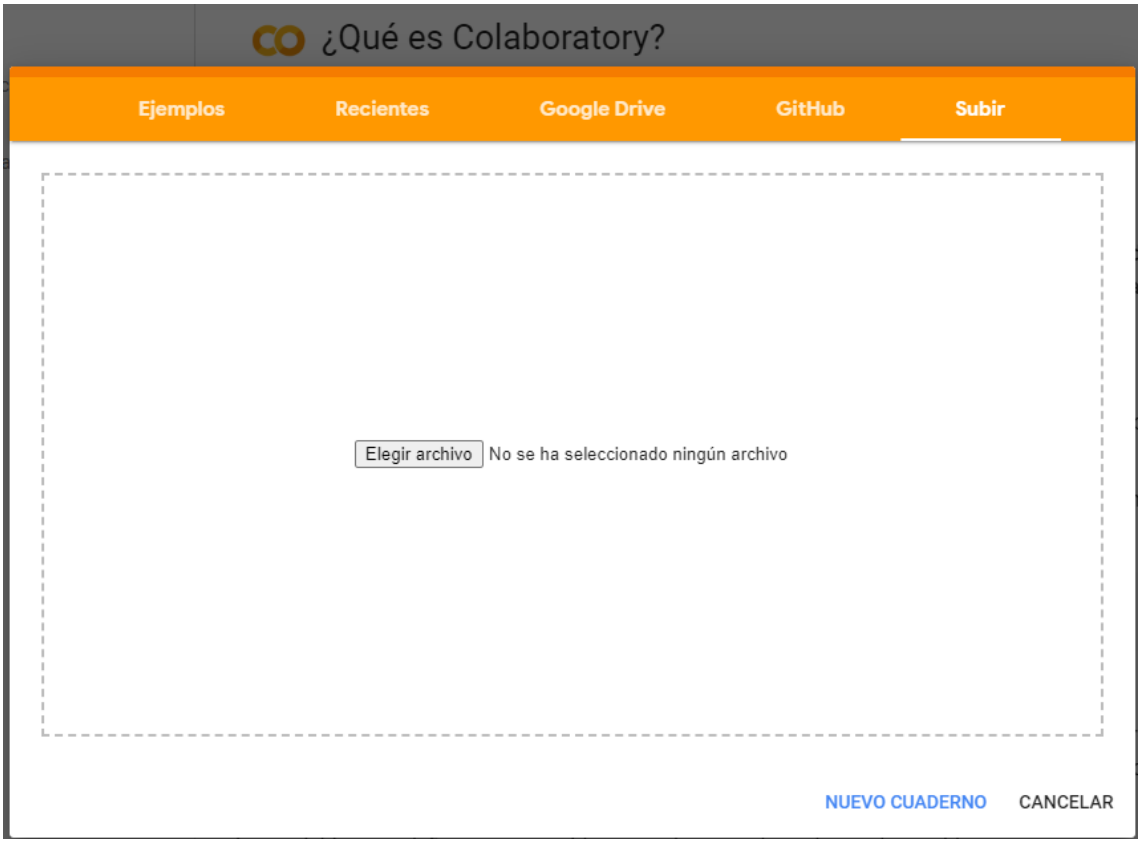
1. N=7, R=3, W=1
2. N=7, R=1, W=7
3. N=7, R=5, W=4
4. N=7, R=4, W=5

# Anexo I

Descarga en tu equipo el notebook `IntroMapReduce.ipynb`. El notebook está pensado para que lo ejecutes en Google Colab ( https://colab.research.google.com/ ).
**Google Colab** es un servicio en la nube que permite ejecutar Jupyter Notebooks accediendo con un navegador web. Tiene además las siguientes ventajas:

- Posibilidad de ejecución mediante GPUs
- Basado en Jupyter Notebook pudiendo crear y ejecutar libros en Python 2 o 3
- Tiene preinstaladas las librerías comunes usadas en ciencia de datos y la posibilidad de instalar otras.
- Enlaza con cuentas de Google Drive y desde github

Primero hay que entrar en sesión (login) con una cuenta de Google (la de la uoc debería funcionar). Ahora ya se puede subir el notebook de la PEC a la plataforma:




La ejecución del libro es exactamente igual que en cualquier Jupyter Notebook. Hay que pulsar Shift + Enter para que el código (python) se ejecute. Simplemente vaya ejecutando cada celda y vea el resultado de la ejecución del código.

# 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_pec1.docx. La fecha límite para entregar la PEC2 es el **17 de noviembre.**

# 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.