# Prepara tu examen de SBD

## Introducción

Cada centro, cada año y cada docente, puede plantear al alumnado un modelo de examen concreto que, a su criterio, pueda servir como una correcta evaluación del módulo.

Para ayudar a preparar las evaluaciones, he pensado que podría ser de ayuda crear un archivo único para cada módulo, que pueda crecer cada año con el feedback y apoyo de la comunidad, con cuestionarios de todo tipo, con solucionario o solo los enunciados, pues la intención primera es poder ofrecer una idea de lo que podemos encontrarnos a la hora de una evaluación, poder aprender con ello, y no algo que una persona acabe memorizando, y esperando, sin comprender ni ahondar en la materia, que aparezca mágicamente en el examen.

Este documento, por tanto, no pretende ser una guía única y veraz de exámenes pasados o futuros, pero si una fuente de información sobre la que basar vuestros estudios.

## Posibles modelos.

## Modelo 1.

## Ejemplo pasado por alumnado.

1. **(Puntuación: 2). Defina y describa las 5 características que se usan en Big Data.**
   **Volumen:** Como la gran cantidad de bytes que componen los datos.
   No existe tamaño concreto para los entornos de Big Data, pero hoy día trabajan con volúmenes del orden de los petabytes (PB) e incluso exabytes (EB).
   Son datos abiertos, accesibles, fiables, estructurados, documentados y reutilizables.
   **Velocidad:** El problema no es solo el hecho de que el volumen de datos continúe creciendo sin parar, sino lo rápido que es necesario obtenerlos e integrarlos.
   De la gran velocidad, nacen las estrategias de tipo Streaming.
   **Variedad**: En cuanto a la representación de la información.
   Datos estructurados: registros de bases de datos, con esquema definido y tipo de dato.
   Datos no estructurados: sin esquema, como audios o vídeos, que son el 80% del total.
   Datos semiestructurados: con cierta estructura, pero sin naturaleza relacional (no están en una tabla), almacenados en ficheros con formatos tipo CSV, JSON, XML...
   Metadatos: datos extra muchas veces automáticos para una interpretación posterior.
   **Veracidad:** Los datos no siempre cuentan con la calidad deseada o no son fieles a la realidad. Está relacionado con los conceptos de señal/ruido.
   Ruido son los datos que no pueden ser convertidos en información, porque no contienen o está corrupta.
   Señal son datos que pueden ser convertidos en información con sentido.
   Por ello es necesario conocer en qué condiciones se adquirieron los datos y procesarlos.
   Los datos obtenidos en modo automático (como crear transacciones) contienen menos ruido que los de personas (post de un blog).

**Valor**: Cómo de útiles son. Es importante su veracidad, el tiempo transcurrido, que sean completos o su interpretación.
Etapas desde los eventos al valor:
Los **eventos** se producen, se generan **datos** que pueden ser almacenados y dando contexto tenemos **información** para tras darles significado obtener **conocimiento**. Con él, creamos modelos predictivos para obtener **sabiduría**, con la que realizar acciones que generen **valor**.

2. **(Puntuación: 1). Describa los conceptos de *Sharding* y Replicación.**
**Sharding:** dividir los conjuntos de datos en subconjuntos (shards) para distribuirlos por el clúster. Facilita tratar con datos más grandes, la distribución entre nodos para una escalabilidad horizontal y proporciona tolerancia parcial a fallos.
El problema es que debe diseñarse la estrategia de sharding (el algoritmo o fórmula que decide qué registro va a cada shard), según las necesidades de acceso/consulta de datos.
**Replicación:** Hacer copias de los datos en distintos nodos para una mayor disponibilidad y tolerancia a fallos. Hay dos estrategias:
Maestro-esclavo (master-slave): Donde todas las escrituras son en el nodo maestro y después son replicadas. El maestro es el punto único de fallo. Las lecturas son en cualquier nodo. Por ello, está indicado para lectura intensiva (al usar todos los nodos), pero no para escritura intensiva (al ser solo el nodo maestro).
Par-a-par (peer-to-peer): No hay nodo maestro y todos (peers) están al mismo nivel jerárquico. Por esto, no hay punto único de fallo. Puede escribir y leer en todos los nodos. Se escribe en uno y se replica posteriormente.
Puede haber inconsistencias de escritura. La concurrencia se gestiona de forma:
Pesimista, asegurando que no se puede modificar a la vez en dos nodos distintos.
Optimista, donde no hay bloqueos y la base de datos está siempre disponible.
**Sharding con replicación:** Combina las dos para obtener las ventajas de ambas.
Sharding con replicación maestro-esclavo, donde en cada shard hay un maestro y *N* nodos esclavos. El maestro es el punto único de fallo y las escrituras afectadas a ese shard se realizan en el maestro y se propagan.
Sharding con replicación par-a-par: En cada shard hay réplicas al mismo nivel de jerarquía. No hay punto único de fallo y las lecturas y escrituras pueden ser en cualquiera.

3. **(Puntuación: 1). Definición de RAID y niveles más empleados.**
Sistema de almacenamiento de datos distribuidos por múltiples unidades, según estrategias o niveles. No es un sistema de ficheros sino una capa por debajo de ellos. Puede trabajar por hardware mediante controladoras RAID o por software, donde el procesador ejecuta un software equivalente a la controladora hardware (más lento pero más barato). Los niveles son:
**RAID 0:** Conjunto dividido. Dos discos de 1 TB como uno de 2 TB sin replicación. Permite unidades virtuales más grandes, sin redundancia, a doble velocidad en lectura y escritura.
**RAID 1:** Conjunto en espejo. Dos discos de 1 TB como uno de 1TB con replicación. Datos duplicados de forma automática. Mayor seguridad desaprovechando espacio. Hasta el doble de velocidad en lectura y tolerante al fallo en uno de los discos.
**RAID 5:** Conjunto dividido con paridad distribuida. Tres discos de 1 TB como uno de 2 TB con replicación. Mínimo tres discos. Unidades virtuales más grandes que las

físicas añadiendo replicación. Mayores rendimientos de lectura y tolerante al fallo en uno de los discos.

4. **(Puntuación: 2). Describa las distintas fases en un proceso ETL.**
   **Extraer:** se adquieren o se extraen los datos incluyendo una validación previa. Suelen ser fuentes compatibles con ETL. Es importante elegir bien la estrategia para recibir datos, sobre todo en sistemas transaccionales.
   **Transformar:** Transformación de los datos en el formato deseado. Incluye limpiar datos y crear datos sintéticos. Emplea reglas para prepararlos, asegurando que son válidos y que tienen el formato deseado. Es el proceso que mayor esfuerzo humano requiere (seleccionar sólo unas columnas o atributos, eliminar datos duplicados, ordenar, traducir valores…).
   **Cargar:** Carga los datos en destino. Normalmente un almacén de datos para uso OLAP, un sistema de ficheros distribuido (HDFS o Amazon S3), ficheros planos o sistemas OLTP (migrar ERP). Fase sencilla y liviana.
   Podemos cargar los datos completos, solo los datos nuevos, o cargar de forma incremental (puede implicar actualizar los cargados).
5. **(Puntuación: 2). Describa los distintos niveles de Análisis en Analítica de Datos.**
   **Análisis Descriptivo:** ¿qué ha ocurrido? Intenta descubrirlo con un cuadro de mando estático con consultas operacionales.
   **Análisis Diagnóstico:** ¿por qué ha ocurrido? Intenta determinar la causa de un fenómeno, mostrando herramientas interactivas para identificar tendencias y patrones.
   **Análisis Predictivo:** ¿qué va a ocurrir? Utiliza modelos predictivos de ML.
   **Análisis Prescriptivo:** ¿qué hacer para que ocurra? Se apoya en los resultados producidos por el Análisis Predictivo.
6. **(Puntuación: 1). Características diferenciales de cuadro de mandos e informe.**
   Los **informes** están más enfocados a instantánea que al análisis en tiempo real. Más enfocados a página y a tablas. Ocupan más espacio, tienen mayor cantidad de datos históricos y suelen enviarse de forma periódica.
   Los **cuadros de mando** suelen ser más interactivos y dinámicos, con cambios en tiempo real. Más enfocados a gráficos y pantalla. Más resumidos, intentando mostrar con un rápido vistazo. Pueden mostrar alarmas sobre métricas.
7. **(Puntuación: 1). Datawharehouse y Datamart, qué son y como se pueden combinar.**
   Los **Datawharehouse** o almacén de datos, son repositorios que guardan tanto información histórica como reciente. Una instantánea de información denormalizada, almacenada para un alto rendimiento en analítica.
   Los **Datamart** son un subconjunto de la información de un departamento o división. Si están con un Datawharehouse, serán datamarts dependientes.
   Los **Datamarts independientes** tienen la ventaja de eliminar el Datawharehouse central y disminuir la complejidad del sistema, pero no puede realizar informes o cuadros de mando transversales, y deja de tener una versión única.

## Modelo 2.

## Ejemplo pasado por alumnado.

1. (Puntuación:1). Diferencias entre Datos, Información y Conocimiento.

2. (Puntuación:2). Enumere y describa brevemente las distintas capas usadas en una arquitectura de Big Data.
3. (Puntuación:1). Tipos de bases de datos NoSQL.
4. (Puntuación:1). HDFS, características y funcionamiento.
5. (Puntuación:1). Características y diferencias de ETL e Integración de Datos.
6. (Puntuación:1). Enumere las principales metodologías en Minería de Datos, describiendo las principales tareas y fases.
7. (Puntuación:1). Características deseables de una métrica de un cuadro de Mandos.
8. (Puntuación:1). Tipos de Inteligencia de Negocio.