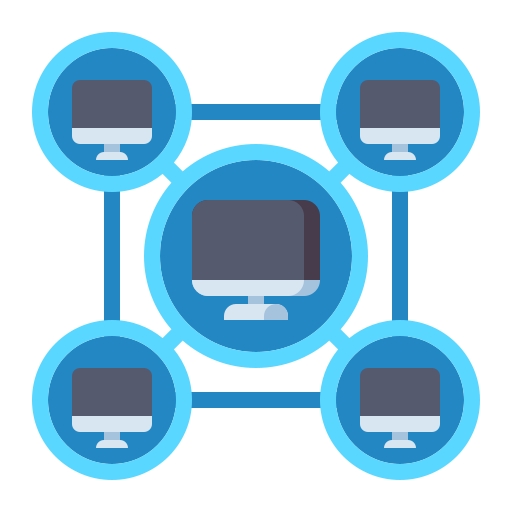
**Big data aplicado**

**Caso práctico**

**Construcciones d8**

**Índice**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

[**1. Enunciado 2**](#_2mb3c51ofwfw)

[**2. Apartado 1 3**](#_e3bkkjfrb8sm)

[**3. Que se evaluará de la tarea 3**](#_dupk3rs459rh)

[**3.1 ¿Cómo debería ser su escalabilidad? 3**](#_83n3lqn3ekb6)

[**3.2 ¿Qué capas son necesarias como mínimo? 3**](#_cdlyg8bqs27o)

[**3.3 ¿Deberá cumplir las características ACID? 4**](#_lcsk0fyiistt)

[**3.4 ¿Será OLTP? 4**](#_r6u0wgef8dq7)

[**3.5 ¿Será OLAP? 4**](#_k1y6lg9m7cry)

[**3.6 ¿Necesitará un almacén de datos? 4**](#_i82a3cdxm8by)

[**3.7 ¿Qué estrategia debería seguir? 4**](#_8zdbhxe7mf6b)

[**3.8 ¿Es necesario crear modelos predictivos? 5**](#_74zgxrsj7ui)

## **1. Enunciado**

La empresa Construcciones D8 se ha puesto en contacto con la empresa consultora en la

que trabajas para que le realices un prediseño de lo que sería un sistema Big Data para

resolver las siguientes necesidades.

● Hay distintas fuentes externas a su empresa que producen datos interesantes para

ellos y les interesaría poder conectarse a ellas para obtenerlos.

● Esas fuentes tienen conjuntos de datos estáticos o que se actualizan anualmente.

● Además hay fuentes internas de la propia empresa que generan datos de forma

contínua y hay que irlos obteniendo sobre la marcha.

● La cantidad de datos actualmente es de aproximadamente 500TB, y calculan que se

producen otros 100TB nuevos cada año.

● Quieren poder mantener almacenados todos esos datos de modo no se pierdan y

además accesibles en todo momento.

● Se realizan transacciones debido a la interacción con clientes en el día a día.

● La junta directiva se reúne una vez al mes y quiere poder acceder a un cuadro de

mandos para ver analíticas descriptivas que empleen todos los datos que estuviesen

disponibles una semana antes de reunirse. Tales analíticas deben ser interactivas,

siendo los directivos capaces de realizar filtrados de información de modo que las

gráficas mostradas se actualicen según la información seleccionada.

● Quieren poder decidir a qué clientes ofrecerles ciertas ofertas en función de lo que

se sabe de su comportamiento pasado.

## **2. Apartado 1**

**Prediseño un sistema para Big Data**

Crea un documento en el que explicas cómo sería el sistema a emplear para resolver las

necesidades Big Data del supuesto práctico. Deberás:

● Indicar qué habrá que hacer para ir aumentando la capacidad del cluster según se

reciben nuevos datos.

● Indicar qué capas de la arquitectura Big Data necesitarán estar presentes como

mínimo en el sistema a crear.

● Indicar si alguna parte del sistema necesitará cumplir con las características

ACID.(Atomicidad, Consistencia, Aislamiento, Durabilidad)

● Indicar si será necesario un subsistema OLTP.

● Indicar si será necesario un subsistema OLAP.

● Indicar si habrá un almacén de datos.

● Indicar qué estrategia de procesamiento habrá que emplear para poder crear el

cuadro de mandos que quiere la junta directiva.

● Indicar si será necesario crear modelos predictivos a partir de los datos.

## **3. Que se evaluará de la tarea**

### **3.1 ¿Cómo debería ser su escalabilidad?**

**Indica correctamente qué habrá que hacer para ir aumentando la capacidad del**

**cluster según se reciben nuevos datos.**

Debería optar por una escalabilidad horizontal. La razón principal es que están manejando un conjunto de datos que crece rápidamente, requieren procesamiento en tiempo real y acceso constante a grandes volúmenes de datos, todo lo cual se adapta mejor a un enfoque de clúster distribuido. Además, este enfoque facilitará el crecimiento futuro sin encontrarse con limitaciones físicas a medida que se incremente la carga de trabajo y el volumen de datos.

### **3.2 ¿Qué capas son necesarias como mínimo?**

**Indica correctamente qué capas de la arquitectura Big Data necesitarán estar**

**presentes como mínimo en el sistema a crear.**

Como mínimo, deberá tener presente la arquitectura de dos capas, ya que necesitamos que los datos estén bien separados entre ellos, además de que nos permite un mejor acceso a los datos históricos y un mejor análisis de los datos que en la arquitectura de una capa.  
  
Pero a largo plazo daría problemas ya que es menos efectiva para garantizar la calidad y escalabilidad de los datos. Por lo tanto la mínima es de dos capas pero la óptima sería de tres capas.

### **3.3 ¿Deberá cumplir las características ACID?**

**Indica correctamente si alguna parte del sistema necesitará cumplir con las**

**características ACID.**

Deberá cumplir todas las características ACID; duradera para evitar la pérdida de información; atómica garantiza que todos los pasos de una sola transacción de base de datos se completen o se reviertan a su estado original; consciente para mantener la integridad de la base de datos y reglas de la empresa; Aislada para garantizar que las transacciones simultáneas no interfieran entre sí.

### **3.4 ¿Será OLTP?**

**Indica correctamente si será necesario un subsistema OLTP.**

Sí, será necesario un subsistema OLTP para gestionar las transacciones diarias e interacciones con los clientes. Este subsistema garantizará que las operaciones sean rápidas y consistentes.

### **3.5 ¿Será OLAP?**

**Indica correctamente si será necesario un subsistema OLAP.**

También será necesario un subsistema OLAP para realizar análisis complejos sobre grandes volúmenes de datos. Esto permitirá a la junta directiva realizar análisis descriptivos sobre la información almacenada.

### **3.6 ¿Necesitará un almacén de datos?**

**Indica correctamente si habrá un almacén de datos.**

Deberá implementar un Data Warehouse para almacenar los datos históricos, facilitar su análisis y acceso de cara a la junta directiva y al guardado y consulta de la información.

### **3.7 ¿Qué estrategia debería seguir?**

**Indica correctamente qué estrategia de procesamiento habrá que emplear para**

**poder crear el cuadro de mandos que quiere la junta directiva.**

Para el cuadro de mandos interactivo solicitado por la junta directiva, es necesario emplear una estrategia de procesamiento que combine el procesamiento en; Batch para datos que no necesitan ser analizados en tiempo real (como los datos estáticos actualizados anualmente), se pueden programar cargas de datos regulares y realizar análisis en lotes usando Apache Spark o Hadoop; y procesamiento en streaming: Los datos generados de manera continua (por ejemplo, de las interacciones diarias con clientes) deben ser procesados en tiempo real mediante herramientas como Apache Kafka o Apache Flink para que los análisis sean lo más actuales posible.

### **3.8 ¿Es necesario crear modelos predictivos?**

**Indica correctamente si será necesario crear modelos predictivos a partir de los**

**datos**

Será necesario crear modelos predictivos a partir de los datos recopilados. Esto permitirá a la empresa ofrecer ofertas personalizadas a los clientes basadas en su comportamiento pasado. Para esto, se podrán utilizar herramientas de machine learning como Scikit-Learn, TensorFlow o PyTorch.