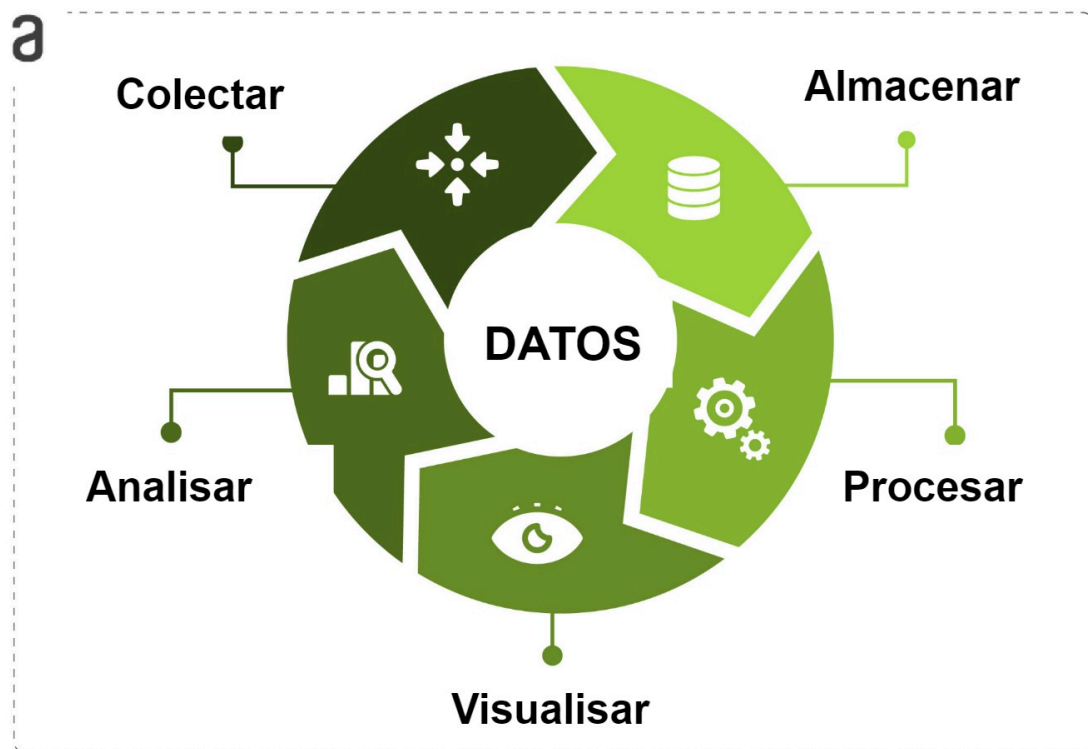


“Del dato bruto a la decisión”



Adrián Yared Armas de la Nuez

Contenido

| | |
|--|----------|
| 1. Contexto..... | 2 |
| 2. Verificación de calidad e integridad de los datos..... | 2 |
| 2.1 Identificar al menos dos problemas adicionales..... | 2 |
| 2.1.1 Problema 1..... | 2 |
| 2.1.2 Problema 2..... | 2 |
| 3. Diseño de almacenamiento escalable..... | 3 |
| 3.1 Gestión de datos no estructurados..... | 3 |
| 3.1.1 contexto, estructura general..... | 3 |
| 3.1.2 Contexto, formato de almacenamiento por tipo de dato..... | 4 |
| 3.1.3 Gestión de datos no estructurados..... | 4 |
| 3.2 Justificar las decisiones técnicas tomadas..... | 5 |
| 3.3 Modelado de KPIs y visualización..... | 5 |
| 3.3.1 KPIs propuestos..... | 5 |
| 3.4 Simulación de resultados y validación..... | 6 |
| 3.4.1 Inventar un escenario de simulación..... | 6 |
| 3.4.2 Proponer decisiones asociadas..... | 7 |
| 3.4.3 Indicar cómo validarías el modelo..... | 7 |
| 3.5.1 ¿Qué consecuencias tiene usar datos sin validar en decisiones de negocio?..... | 7 |
| 3.5.2 ¿Cómo cambia el análisis cuando los datos son no estructurados?..... | 8 |
| 3.5.3 ¿Qué riesgos éticos existen en el uso de datos personales?..... | 8 |
| 3.5.4 ¿Qué parte del proceso de BI has encontrado más desafiante y por qué?..... | 8 |

1. Contexto

La empresa SmartRetail, dedicada al comercio electrónico, busca implementar un sistema de inteligencia de negocios (BI) que le permita transformar datos masivos en decisiones eficientes. Cuentan con datos estructurados (ventas, clientes) y no estructurados (opiniones en redes sociales) almacenados en un sistema distribuido. Te han contratado como especialista en Big Data para validar los datos, diseñar la arquitectura de almacenamiento, definir indicadores de negocio (KPIs), simular resultados e interpretar decisiones estratégicas.

2. Verificación de calidad e integridad de los datos

Enunciado: Identifica problemas comunes en la calidad e integridad de datos y propone mecanismos de detección y corrección.

Ejemplo desarrollado:

- Problema: Formatos de fecha no homogéneos.
- Solución: Estandarización con funciones de PySpark.
- Validación: Aplicación de hashes para detectar cambios.

2.1 Identificar al menos dos problemas adicionales.

2.1.1 Problema 1

Como primer problema podemos encontrar el mal uso de los datos campos del registro de ventas, pudiendo estos estar nulos o incompletos, como por ejemplo, productos faltantes, datos del cliente o el id del cliente.

Una posible solución podría ser el uso de `na.drop()` de PySpark para eliminar registros incompletos si son imposibles de reparar o el uso de `na.fill()` para rellenar los valores de una manera estándar o estimada según consideres.

En cuanto a la validación, realizaría un análisis exploratorio con `df.describe()` y `df.select()` para detectar valores nulos.

Es decir, ante la problemática del mal uso de los datos, utilizaría PySpark para rellenar o eliminar los registros según considere en cada caso.

2.1.2 Problema 2

Como segundo posible problema, esta vez por parte de los datos no estructurados recibidos de los comentarios en redes sociales, tendríamos el ruido y uso coloquial de las redes sociales, tales como: sarcasmo; uso de emojis; spam; lenguaje ofensivo. Ya que para el humano podemos interpretar el significado de un sarcasmo o significado de un emoji, pero para una máquina es algo más complicado.

Como posible solución propongo el uso de herramientas como llms, algoritmos de análisis de sentimientos o herramientas de procesamiento de lenguaje natural, con esto podríamos tener ideas como la aceptación o rechazo del producto con el análisis de sentimientos o un análisis algo más profundo con las aplicaciones de procesamiento de lenguaje natural, ya que permitiría la eliminación de emojis, símbolos no relevantes, y palabras ofensivas. Además utilizaría detecciones de repetición excesiva para evitar el spam y la detección de bots publicitarios.

Para asegurar la calidad tras limpiar las opiniones en redes sociales, analizaría la longitud de los textos antes y después, observando si se eliminó ruido sin cortar contenido útil, además analizaría la frecuencia de palabras clave (como por ejemplo "envío") para confirmar que la limpieza no borra términos importantes, además intentaría medir el porcentaje de textos descartados por spam o lenguaje ofensivo, manteniéndolo dentro de un rango razonable, ya que en muchos casos el contexto social y otros factores pueden normalizar el uso de lenguaje soez y puede ser comentario de valor para la empresa. Y finalmente, utilizaría gráficos de barras u otro tipo de gráficas para revisar los cambios de forma clara y rápida.

3. Diseño de almacenamiento escalable

Enunciado: Diseña una arquitectura de almacenamiento adecuada para SmartRetail.

Ejemplo desarrollado:

- Estructura: Almacenamiento HDFS con carpetas por categoría de datos.
- Formato: Parquet para eficiencia y compresión.

Recurso de apoyo:

<https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html>

3.1 Gestión de datos no estructurados

3.1.1 contexto, estructura general

Utilizaría un HDFS como sistema de almacenamiento principal debido a su escalabilidad, tolerancia a fallos y capacidad para manejar grandes volúmenes de datos.

Y tendría una estructura de carpetas similar al ejemplo dado:

```
/smartretail/  
├── ventas/  
│   ├── 2025/  
│   └── 2024/  
├── clientes/  
├── productos/  
├── redes_sociales/  
│   ├── twitter/  
│   ├── facebook/  
│   ├── instagram/  
│   └── .../  
└── reportes/
```

3.1.2 Contexto, formato de almacenamiento por tipo de dato

| Tipo de dato | Formato elegido | Justificación |
|---------------------------------|-----------------|--|
| Datos estructurados | Parquet | Formato columnar, comprimido y optimizado para consultas en Spark y Hive. |
| Datos no estructurados (texto) | JSON | Formato flexible, ideal para opiniones en redes sociales y fácil de parsear. |
| Datos semi-estructurados (logs) | Avro | Ideal para flujos de datos con esquemas evolutivos y alto rendimiento. |

3.1.3 Gestión de datos no estructurados

Para gestionar los datos no estructurados, como las opiniones de redes sociales, se plantea una estrategia eficiente que permite capturar, almacenar y analizar esta información en tiempo real. Para ello, se utilizan herramientas como Apache Flume o Kafka para automatizar la inserción de datos desde las APIs de redes sociales, capturando menciones y comentarios conforme se publican. Una vez recolectados, estos datos se almacenarán en HDFS utilizando el formato JSON. La organización se hará por red social y fecha, lo que facilita el acceso y la trazabilidad.

Por ejemplo:

```
/smartretail/redes_sociales/twitter/2025/05/20/opinion.json
```

Posteriormente, se procesan con PySpark y técnicas de procesamiento de lenguaje natural para llevar a cabo tareas como el análisis de sentimiento, la detección de spam o la

extracción de términos clave y entidades mencionadas. y finalmente, si se requiere realizar búsquedas rápidas o visualizaciones basadas en texto, los resultados podrán indexarse en una base de datos, permitiendo una exploración más ágil de las conversaciones relevantes para el negocio.

3.2 Justificar las decisiones técnicas tomadas.

Por la parte técnica he decidido utilizar un hdfs, ya que permite distribuir el almacenamiento a bajo coste y con alta disponibilidad, esencial para manejar tanto ventas históricas como flujos en tiempo real. También he decidido utilizar parquet, ya que aumenta la eficiencia de las lecturas selectivas de columnas. También decidí utilizar json para redes sociales, ya que permite almacenar estructuras de datos de manera flexible, como el caso de los comentarios. Y finalmente decidí separarlas por tipo y año para facilitar las consultas y simplificar el mantenimiento de datos históricos.

3.3 Modelado de KPIs y visualización

3.3.1 KPIs propuestos

Como primer KPI se me ocurre introducir el sentimiento promedio de las opiniones en redes sociales por categoría de producto. Este permite evaluar la percepción de los clientes en tiempo real, útil para ajustar campañas o detectar problemas en productos específicos. Por parte de la visualización, utilizaría mapas de calor por categoría vs. sentimiento (positivo, neutro, negativo), con el fin de ver de manera gráfica y sencilla la visualización de la aceptación de cada producto dentro del subgrupo de productos al que pertenece. Como herramienta de BI utilizaría Tableau, ya que es excelente para conexión con fuentes no estructuradas como JSON.

Como segundo KPI, introduciría el tiempo promedio desde la visita al sitio web hasta la compra, lo cual evaluaría la eficiencia de la web, productos y recomendaciones al usuario. Ya que un tiempo muy largo puede indicar resistencia o incomodidad a la hora de realizar la compra.

Por la parte de la visualización utilizaría gráficos de barras apiladas por dispositivo (móvil, desktop, tablet). Para conocer y adaptarme con mayor facilidad a las necesidades del usuario.

Como herramienta de BI utilizaría Google Looker Studio (ideal para análisis de comportamiento web).

El tercer BI que se me ocurre es la tasa de abandono del carrito de compras y que mida el porcentaje de usuarios que agregan productos al carrito pero no completan la compra, ya que es un indicador clave para detectar problemas en la experiencia de usuario o en el proceso de pago.

Por parte de la visualización haría un gráfico de embudo mostrando las etapas del proceso de compra y el resultado del cálculo del abandono del carrito. Y como herramienta de BI utilizaría Power BI (por su capacidad para construir embudos interactivos y aplicar filtros).

3.4 Simulación de resultados y validación

Enunciado: Simula datos y plantea decisiones de negocio en base a los resultados obtenidos.

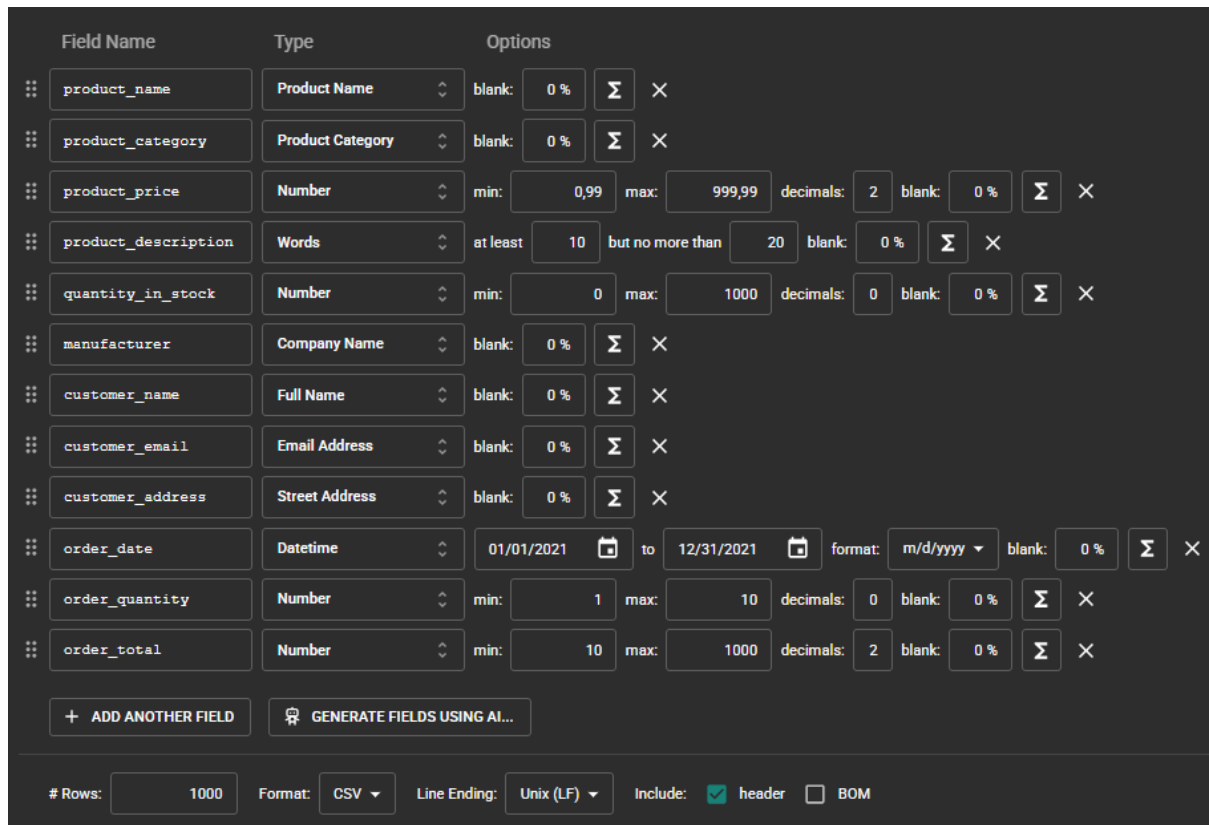
Ejemplo desarrollado:

- Simulación: Segmento 25-34 años genera 60% de compras online.
- Decisión: Campaña de fidelización digital.
- Validación: Comparación con datos históricos.

Recurso de apoyo: <https://www.mockaroo.com/>

3.4.1 Inventar un escenario de simulación

Creé un escenario de simulación a través de su herramienta gráfica con campos referidos al usuario, producto y compras.



| Field Name | Type | Options |
|---------------------|------------------|--|
| product_name | Product Name | blank: 0 % |
| product_category | Product Category | blank: 0 % |
| product_price | Number | min: 0,99 max: 999,99 decimals: 2 blank: 0 % |
| product_description | Words | at least 10 but no more than 20 blank: 0 % |
| quantity_in_stock | Number | min: 0 max: 1000 decimals: 0 blank: 0 % |
| manufacturer | Company Name | blank: 0 % |
| customer_name | Full Name | blank: 0 % |
| customer_email | Email Address | blank: 0 % |
| customer_address | Street Address | blank: 0 % |
| order_date | Datetime | 01/01/2021 to 12/31/2021 format: m/d/yyyy blank: 0 % |
| order_quantity | Number | min: 1 max: 10 decimals: 0 blank: 0 % |
| order_total | Number | min: 10 max: 1000 decimals: 2 blank: 0 % |

Rows: 1000 Format: CSV Line Ending: Unix (LF) Include: ☒ header ☐ BOM

Y el csv resultante tiene la siguiente estructura:

| product name | product category | product price | product description | quantity in stock | manufacturer |
|-------------------|-------------------------|-----------------------|------------------------------|-------------------|--------------|
| Vegan Mayonnaise | Food - Condiments | 542.25 | ante vel ipsum praesent b | 240 | Babbleopia |
| Coconut Cream | Food - Baking & Cooking | 330.0 | eu tincidunt in leo maecen | 682 | Topiclounge |
| Scent Diffuser | Home | 523.31 | laoreet ut rhoncus aliquet | 558 | Skilith |
| Foot Massader | Health | 577.07 | orci luctus et ultrices posu | 643 | Abata |
| customer name | customer_email | customer_address | order date | order quantity | order total |
| Augy Burwell | aburwell0@vinaora.com | 83394 Butternut Plaza | 3/31/2021 | | 5740.28 |
| Bernita Skitterel | bskitterel1@pen.io | 6 8th Pass | 6/7/2021 | | 1503.56 |
| Cherilynn Petters | cpetters2@usnews.com | 02 Amoth Terrace | 8/25/2021 | | 6733.05 |
| Adamo Dingle | adinale3@opera.com | 0753 Brown Pass | 3/17/2021 | | 632.37 |

3.4.2 Proponer decisiones asociadas.

Como cosas a tener en cuenta en función a los datos registrados encuentro:

El control de consumo de productos por categoría, lo que permite saber qué es lo que más se vende (referido a salud, alimentación, etc) y así mover el stock en función a la prioridad por área de consumo.

El control de stock por consumo, en función al consumo permite hacer una demanda de stock ajustada a la tendencia de las necesidades.

También, en función de las tendencias de compras de usuarios en función al histórico podrías patrocinar productos relacionados de manera interactiva como anuncios flash o con bajo stock.

También podría realizar ajustes de precios en función a la demanda, oferta y consumición

Y podría patrocinar o incluso dejar de ofertar y comprar stock de productos en función a las compras por tipo de artículo en función a la marca, como podría ser comprar mayonesa, en el caso de que tengas dos marcas y una se venda mucho, podrías patrocinarla, y en el caso de otra que no se venda, podrías hacer un estudio de rentabilidad y quizás dejar de comprar stock.

Finalmente, con respecto al KPI anunciado previamente, podría implementar un sistema de recordatorio de carrito abandonado con mensajes personalizados y descuentos progresivos (por ejemplo, 10% si no compró en 24 h).

3.4.3 Indicar cómo validarías el modelo.

Para validar las decisiones tomadas, utilizaría varios enfoques. Primero, se comprobaría los resultados con datos históricos de campañas similares para detectar patrones de abandono. Luego, aplicaría una prueba por segmentación de público, enviando la promoción solo a una parte del público para evaluar su impacto antes de ampliarla.

También se analizaría el recorrido del usuario con herramientas de analítica web, identificando en qué etapa exacta del proceso ocurre el abandono del carrito. Y por último, monitorizaría las redes sociales para detectar cambios en el tono de los comentarios, especialmente en relación a precios y promociones. Esta combinación permitirá medir la efectividad y ajustar las acciones si es necesario.

3.5 Reflexión crítica

Responde con argumentación razonada a las preguntas:

3.5.1 ¿Qué consecuencias tiene usar datos sin validar en decisiones de negocio?

Utilizar datos sin validar en las decisiones puede ocasionar decisiones erróneas y costosas. Si los datos contienen errores, duplicaciones o valores faltantes, los análisis resultantes serán imprecisos, lo que puede generar ruido en el mercado o del comportamiento del cliente. Por ejemplo, un error en los precios puede llevar a mala reputación para la empresa y un desajuste en las necesidades del cliente.

3.5.2 ¿Cómo cambia el análisis cuando los datos son no estructurados?

El análisis de datos no estructurados, a diferencia de los datos estructurados, que se organizan en tablas y se procesan fácilmente con SQL, necesitan técnicas como procesamiento de lenguaje natural, reconocimiento de patrones, minería de datos o machine learning. Esto no solo aumenta la complejidad técnica y tratamiento del dato, sino que también introduce retos en interpretación, limpieza y contextualización de los datos. Sin embargo, pese a su complejidad de tratamiento, interpretación y uso, son datos muy valiosos para las empresas.

3.5.3 ¿Qué riesgos éticos existen en el uso de datos personales?

El uso de datos personales conlleva responsabilidades éticas. Como en el caso de la invasión de la privacidad, la recopilación “sin consentimiento” de los datos personales, la discriminación algorítmica y la filtración de información sensible.

Por ejemplo, un mal manejo de datos de clientes puede derivar en violaciones de normativas de seguridad del uso de los datos, como el caso de la Ley de protección de datos española. Además, cuando se utilizan datos para segmentar o personalizar ofertas, hay que evitar prácticas manipulativas o sesgos que afecten a ciertos grupos sociales.

3.5.4 ¿Qué parte del proceso de BI has encontrado más desafiante y por qué?

La parte más desafiante del proceso de BI para mí fue la interpretación de los datos, ya que al no ser experto en el dato me ha costado algo de tiempo analizar y comprender el verdadero alcance de la información y la potencia de los datos y la correlación entre ellos. Pero pese a la dificultad, el BI me parece una herramienta extremadamente útil tanto para la implementación en el mercado como para aprender a interpretarlo cuando lo encuentras en el día a día, y así evitar manipulaciones y malos usos de la ética algorítmica.