“Del dato bruto a la decisión”

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.

Tarea 1: Verificación de calidad e integridad de los datos (20 min)

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.

- Identificar al menos dos problemas adicionales.

- Problema: Los datos pueden contener valores incorrectos, mal estructurados o poco fiables, afectando su interpretación.
- Solución: Aplicación de reglas de validación para detectar valores atípicos y mecanismos de corrección para normalizar formatos, asegurando coherencia.
- Validación: Aplicación de reglas de integridad que detecten valores fuera de rango o formatos incorrectos, complementado con auditorías de datos para garantizar coherencia.

- Problema: La repetición de registros puede distorsionar análisis y generar información errónea.
- Solución: Implementación de técnicas de detección y eliminación de duplicados, utilizando reglas de negocio para consolidar información sin pérdida de datos clave.
- Validación: Uso de técnicas de comparación de registros, como claves únicas o algoritmos de detección de similitud, asegurando la eliminación de duplicados sin afectar la calidad de la información.

- Proponer herramientas concretas para validarlos.

- Inconsistencias en formatos y valores

PySpark DataFrame API: Uso de funciones como regexp_extract() para estandarización y limpieza.

- Datos duplicados

PySpark dropDuplicates(): Filtra registros repetidos basándose en columnas clave.

Tarea 2: Diseño de almacenamiento escalable (15 min)

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

- Describir cómo gestionarías los datos no estructurados.
- Justificar las decisiones técnicas tomadas.

Estructura:
- Almacenamiento distribuido en HDFS, con carpetas organizadas por tipo de datos (ventas, clientes, redes sociales).

- Uso de bases de datos NoSQL como MongoDB o Cassandra para datos no estructurados, facilitando la consulta y recuperación eficiente.

Formato:
- Parquet para datos estructurados, optimizando la compresión y velocidad de acceso.

- JSON para datos no estructurados, permitiendo flexibilidad en el almacenamiento y procesamiento.

Gestión de datos no estructurados:
- Apache Kafka para la ingesta en tiempo real de datos provenientes de redes sociales.

- Spark Streaming para el procesamiento en tiempo real y análisis de tendencias.

- Elasticsearch para indexación y búsqueda eficiente de datos textuales.

Justificación de decisiones técnicas:

- HDFS: Permite almacenamiento distribuido y escalable, ideal para grandes volúmenes de datos.

- Parquet: Mejora la eficiencia en almacenamiento y consulta de datos estructurados.

- JSON: Facilita la gestión de datos no estructurados con esquemas flexibles.

- Kafka y Spark Streaming: Garantizan procesamiento en tiempo real para datos dinámicos.

- Elasticsearch: Optimiza la búsqueda y recuperación de información clave.

Tarea 3: Modelado de KPIs y visualización (25 min)

Enunciado: Define indicadores clave de negocio y visualizaciones adecuadas.

Ejemplo desarrollado:
- KPI: Ratio de conversión mensual por canal.
- Visualización: Gráfico de líneas en Power BI.

- Proponer tres KPIs adicionales.
- Asociar tipo de visualización y herramienta de BI.

1. Ticket promedio por cliente

- Descripción: Mide el gasto promedio de cada cliente en un período determinado.
- Visualización: Gráfico de barras comparativo por segmento de clientes.
- Herramienta: *Power BI* o *Tableau* para análisis detallado.

2. Tasa de abandono del carrito de compras

- Descripción: Indica el porcentaje de clientes que agregan productos al carrito pero no completan la compra.
- Visualización: Gráfico de embudo para visualizar el proceso de compra.
- Herramienta: *Google Data Studio* o *Looker* para análisis web.

3. Sentimiento de clientes en redes sociales

- **Descripción:** Analiza la percepción de los clientes sobre la marca a partir de comentarios y menciones en redes sociales.
- **Visualización:** Nube de palabras o gráfico de sentimiento (positivo, neutro, negativo).
- **Herramienta:** *Power BI* con integración de *Azure Cognitive Services* para análisis de texto.

Tarea 4: Simulación de resultados y validación (30 min)

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/

- Inventar un escenario de simulación.
- Proponer decisiones asociadas.
- Indicar cómo validarías el modelo.

Simulación:

- Se observa que el segmento de clientes de 18-24 años ha incrementado su participación en compras online, representando 45% del total en el último trimestre.

- Los productos más vendidos en este grupo son tecnología y moda, con un crecimiento del 30% respecto al trimestre anterior.

Decisión de negocio:

- Implementar una campaña de descuentos exclusivos para este segmento, enfocada en productos de tecnología y moda.

- Optimizar la experiencia móvil y mejorar la personalización de recomendaciones en la plataforma de e-commerce.

Validación:

- Comparación con datos históricos para verificar si el crecimiento es sostenido o estacional.

- Análisis de impacto de campañas previas en la conversión de clientes jóvenes.

Tarea 5: Reflexión crítica (15 min)

Enunciado: Responde con argumentación razonada a las preguntas:

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

El uso de datos sin validar puede generar errores en la toma de decisiones, afectando la rentabilidad y eficiencia de una empresa.

- Decisiones erróneas: Datos incorrectos pueden llevar a estrategias equivocadas, como inversiones en mercados no rentables.

- Ineficiencias operativas: Procesos internos pueden verse afectados por información inexacta, aumentando costos y reduciendo productividad.

- Pérdida de confianza: Clientes y socios pueden perder confianza en la empresa si las decisiones basadas en datos incorrectos generan problemas en productos o servicios.

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

El análisis de datos no estructurados requiere técnicas avanzadas debido a su falta de formato predefinido.

- Mayor complejidad: No pueden analizarse con herramientas tradicionales como bases de datos relacionales.

- Uso de procesamiento avanzado: Se emplean técnicas como Natural Language Processing (NLP) y Machine Learning para extraer información útil.

- Almacenamiento flexible: Se utilizan Data Lakes en lugar de bases de datos estructuradas.

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

El manejo de datos personales implica riesgos que pueden afectar la privacidad y seguridad de los individuos.

- Violación de privacidad: Uso indebido de datos puede exponer información sensible sin consentimiento.

- Discriminación algorítmica: Modelos de IA pueden generar sesgos que afecten a ciertos grupos de personas si el dataset no está correctamente preprocesado y limpiado.

- Riesgo de robo de identidad: Datos mal protegidos pueden ser utilizados para fraudes y suplantación de identidad.

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

La parte más complicada ha sido la de la interpretación de los resultados obtenidos, porque no basta con recopilar y procesar datos, es fundamental traducirlos en información útil para la toma de decisiones.