Referencia para la Actividad: “Del dato bruto a la decisión”

Contexto empresarial

Nombre de la empresa: AgroMarket Analytics

Sector: Agricultura y distribución alimentaria

Situación: AgroMarket Analytics se dedica a recopilar datos de producción agrícola, precios de mercados locales y patrones climáticos para ayudar a cooperativas y distribuidores a tomar decisiones logísticas y comerciales.

Actualmente manejan datos:
- Estructurados: rendimientos por hectárea, precios de productos por día y localidad.
- No estructurados: comentarios en foros de agricultores, menciones en redes sociales, reportes en PDF.

Su objetivo es construir un sistema de BI que permita:
1. Predecir cuándo y dónde habrá exceso o escasez de producto.
2. Ajustar precios de compra/venta según estacionalidad.
3. Detectar alertas climáticas o de plagas que afecten a la cadena de suministro.

Desarrollo de la Solución BI

Tarea 1: Verificación de calidad e integridad

Problemas detectados:
- Registros climáticos con valores vacíos o corruptos ("N/A", "--").
- Mismo lote de datos repetido en diferentes fuentes.

Herramientas aplicadas:
- Validación con PySpark (`dropDuplicates()`, `na.drop()`, `filter()`).
- Generación de hash para verificar duplicados entre fuentes.

Tarea 2: Arquitectura de almacenamiento

Sistema elegido: HDFS para almacenamiento distribuido con backup incremental en Amazon S3.

Formato:
- Datos estructurados: Parquet (columnar, comprimido)
- Datos no estructurados: JSON para redes sociales, OCR+PDF2Text para reportes PDF

Estructura:
/agromarket/
├── datos_climaticos/
├── precios_mercado/
├── redes_sociales/
├── reportes_pdf/

Tarea 3: KPIs y visualizaciones

KPI	Visualización	Herramienta
Variación semanal de precios	Gráfico de líneas	Power BI
Porcentaje de productos excedentarios	Mapa por zona geográfica	Power BI
Tiempo medio desde cosecha a venta	Histograma	Superset
Menciones negativas por tipo de cultivo	Nube de palabras	Python (Wordcloud)

Tarea 4: Simulación y validación

Simulación: En mayo, los productos con mayor excedente son papas y cebollas en zonas de medianías de Canarias. Las menciones negativas coinciden con presencia de plagas en ciertas fincas.

Decisiones:
- Lanzar oferta flash para venta mayorista de productos excedentarios.
- Redirigir stocks a zonas con menor producción (norte de Tenerife).

Validación:
- Comparación con patrones del año anterior.
- Cruce de datos de plagas con datos climáticos.

Tarea 5: Reflexión final

Desafío principal: Normalizar e integrar datos no estructurados, especialmente los procedentes de PDF y comentarios.

Riesgo detectado: Si los datos climáticos se reciben con retraso, las decisiones comerciales podrían llegar tarde.

Valor del modelo: Aumenta la capacidad de anticipación y permite tomar decisiones proactivas en vez de reactivas.