**Análisis de las 5V's del Big Data**

En el escenario de MarketTrend S.A., se pueden identificar claramente las 5V's del Big Data:

1. **Volumen:** La empresa recibe una **"gran cantidad de datos que reciben diariamente"**.
2. **Velocidad:** La necesidad de **"identificar tendencias en tiempo real"** demuestra que los datos se generan y deben ser procesados a alta velocidad.
3. **Variedad:** Los datos provienen de múltiples fuentes y formatos: **"redes sociales"** (texto, imágenes), **"registros de compras"** (datos estructurados), **"aplicaciones móviles"** (eventos, logs) y **"encuestas"** (datos semi-estructurados).
4. **Veracidad:** La incapacidad de los sistemas actuales para procesar los datos tiene como desafío el garantizar la consistencia y fiabilidad de la información para la toma de decisiones.
5. **Valor:** El objetivo final del proyecto es **"mejorar la toma de decisiones estratégicas y ofrecer información de valor para sus clientes"**

**Diseño de la Arquitectura Big Data Propuesta**

Se propone una arquitectura **Lake House en la nube**, que combina la flexibilidad de un Data Lake para almacenar todo tipo de datos con la capacidad de análisis estructurado de un Data Warehouse.

**Diagrama de la Arquitectura**

El diagrama muestra diversas fuentes de datos (Redes Sociales, Apps Móviles, Registros de Compras) que alimentan una capa de adquisición. Los datos en tiempo real pasan por Apache Kafka, mientras que los datos en lote van a AWS Glue. Ambos flujos depositan los datos en un Data Lake central en Amazon S3. Apache Spark actúa como el motor de procesamiento unificado que lee S3, transforma los datos y los sirve a una capa de análisis y visualización

**Tecnologías por Capa**

* **Adquisición de Datos:**
  + **Apache Kafka:** Para la ingesta de datos streaming provenientes de RRSS y aplicaciones móviles. Se elige por su alta capacidad de rendimiento y tolerancia a fallos.
* **Almacenamiento Distribuido:**
  + **Amazon S3:** Se utilizará como Data Lake para un almacenamiento centralizado, económico y altamente escalable. Ideal para guardar todos los datos de la empresa en su formato original.
* **Procesamiento Distribuido:**
  + **Apache Spark:** Se propone como el motor de procesamiento unificado. Su capacidad para manejar tanto procesamiento batch como en streaming, junto con su alto rendimiento, lo hacen perfecto para analizar los datos del Data Lake.

**Beneficios Identificados**

1. **Visión 360° del Cliente:** Al unificar todas las fuentes de datos, la empresa podrá construir un perfil completo del comportamiento del consumidor, permitiendo personalización y segmentación mucho más efectivas.
2. **Agilidad y Toma de Decisiones en Tiempo Real:** La nueva arquitectura permitirá analizar las tendencias a medida que ocurren.
3. **Escalabilidad a Futuro:** Una arquitectura basada en la nube con tecnologías como S3 y Spark permite a la empresa crecer sin preocuparse por las limitaciones de la infraestructura. El sistema puede escalar para manejar volúmenes de datos y usuarios cada vez mayores de forma costo-eficiente.

**Reflexión sobre Riesgos y Medidas Sugeridas**

**Riesgos y Desafíos**

* **Complejidad Técnica:** La implementación y mantenimiento de tecnologías de Big Data como Kafka y Spark requiere de personal con habilidades especializadas.
* **Gobernanza de Datos:** Sin una gestión adecuada, un Data Lake puede convertirse en un "Data Swamp" (pantano de datos), un repositorio de datos de baja calidad y sin documentación que dificulta su uso.
* **Costos en la Nube:** Si no se gestionan y monitorean adecuadamente, los costos de los servicios en la nube pueden aumentar de forma inesperada.

**Medidas de Seguridad y Calidad**

* **Seguridad:** Se recomienda implementar un control de acceso basado en roles para asegurar que los empleados solo puedan acceder a los datos que necesitan para su trabajo. Además, se debe aplicar **cifrado de datos** tanto en reposo como en tránsito.
* **Calidad:** Es fundamental establecer un **marco de gobernanza de datos**. Esto incluye la creación de un **catálogo de datos** que documente las fuentes, el linaje y el significado de los datos, así como la implementación de rutinas automáticas de limpieza y validación de la información durante el procesamiento con Spark.