### **Proyecto: Plataforma de Datos en Spotify y el Impacto en el Negocio**

**Objetivo**: Este proyecto tiene como objetivo analizar cómo Spotify ha implementado y mejorado su plataforma de datos utilizando servicios en la nube, enfocándose en la recolección, procesamiento y gestión de datos, así como en herramientas internas como **Lexikon**, que han optimizado la toma de decisiones basada en datos y el descubrimiento de insights.

### **1. Introducción**

* **Spotify**: Es una de las plataformas de streaming de música más grandes del mundo, y procesa **1.4 billones de puntos de datos diarios**. Estos datos provienen de las interacciones de los usuarios, como la reproducción de canciones, clics y comportamiento de la aplicación.
* **Problema**: A medida que la cantidad y complejidad de los datos crecía, Spotify necesitaba una plataforma de datos robusta para gestionar y democratizar la información, optimizando así la toma de decisiones y la innovación dentro de la empresa.

### **2. Servicios en la nube y arquitectura de la plataforma**

* **Migración a Google Cloud (GCP)**: Spotify migró su plataforma a Google Cloud, abandonando su infraestructura on-premise, lo que le permitió manejar un creciente volumen de datos de manera escalable y eficiente.

#### **Recolección de datos**

* La plataforma de **recolección de eventos** de Spotify procesa más de **1 billón de eventos** por día.
* Utiliza componentes como **PubSub queues** y pipelines de anonimización para gestionar la entrega y procesamiento de estos eventos, garantizando la privacidad y seguridad de los datos.
* La recolección de datos es fundamental para:
  + Comprender el contenido más relevante para los usuarios.
  + Mejorar la experiencia de usuario respondiendo directamente a sus interacciones.
  + Ofrecer una experiencia personalizada basada en el comportamiento del usuario.

#### **Procesamiento y gestión de datos**

* Más de **38,000 pipelines** manejan las tareas diarias y por hora para procesar datos.
* Los datos se procesan utilizando herramientas como **BigQuery**, **Flink**, y **Dataflow**, y los pipelines se programan y orquestan a través de operadores personalizados de **Kubernetes (K8)**.
* La **gestión de datos** incluye la implementación de controles de acceso, políticas de retención y seguimiento de la trazabilidad (lineage) para garantizar la integridad de los datos y optimizar costos de almacenamiento.

### **3. Herramientas y optimización del descubrimiento de datos: Lexikon**

* **Lexikon**: Es una herramienta interna que Spotify lanzó en 2017 para mejorar el descubrimiento de datos y facilitar el acceso a los insights generados por los científicos de datos, analistas y otros equipos de la empresa.
* **Problema inicial**: Antes de Lexikon, el crecimiento masivo de datasets en BigQuery, tras la migración a GCP, dificultaba el descubrimiento de datos relevantes, lo que ralentizaba la producción de insights. Los equipos producían investigaciones valiosas que quedaban ocultas para otros grupos que podrían beneficiarse de ellas.

#### **Mejoras clave en Lexikon**

1. **Personalización del descubrimiento de datos**:
   1. **Descubrimiento de baja intención**: Para los usuarios que no saben exactamente qué dataset necesitan, Lexikon ofrece recomendaciones personalizadas de datasets basadas en popularidad, uso reciente y relevancia para los equipos. Esto ha incrementado el uso de recomendaciones en un **20% de los usuarios activos mensuales**.
   2. **Descubrimiento de alta intención**: Para los usuarios que buscan un dataset específico, se mejoró el algoritmo de búsqueda priorizando datasets populares y utilizados con frecuencia. Esto aumentó la confianza de los usuarios en los resultados de búsqueda.
2. **Intercambio de conocimientos entre personas**:
   1. **Mapeo de expertos**: Lexikon ahora permite a los usuarios buscar expertos dentro de la comunidad de insights de Spotify. Esto es útil para nuevos empleados o quienes trabajan en nuevos proyectos y necesitan orientación sobre ciertos datasets.
   2. **Slack Bot de Lexikon**: Integrado en Slack, este bot proporciona un resumen rápido del dataset cuando alguien comparte un enlace en un chat, lo que mejora el intercambio de conocimientos entre equipos y ha aumentado la adopción de Lexikon en un **25%**.
3. **Mejora de la experiencia del usuario con datasets**:
   1. **Estadísticas de uso del esquema**: Los usuarios pueden ver qué campos dentro de un dataset son los más consultados, ayudando a identificar rápidamente los campos más relevantes.
   2. **Consultas recientes**: En lugar de depender de ejemplos estáticos, los usuarios pueden explorar las consultas recientes hechas en un dataset, lo que facilita el acceso a ejemplos actuales y útiles.
   3. **Tablas comúnmente unidas**: Lexikon ahora muestra qué tablas suelen unirse a un dataset específico, facilitando el trabajo de los científicos de datos.

#### **Impacto de Lexikon:**

* **Incremento de la adopción**: Después de implementar estas mejoras, la adopción de Lexikon aumentó del **75% al 95%** entre los científicos de datos, convirtiéndola en una de las herramientas más usadas en Spotify, superando incluso a BigQuery UI y Python en uso.
* **Mayor productividad**: Los usuarios ahora experimentan una reducción significativa en el tiempo necesario para descubrir y empezar a usar datasets, lo que acelera la producción de insights y mejora la toma de decisiones.

### **4. Impacto en el negocio**

* **Acceso democratizado a los datos**: Spotify ha facilitado el acceso a los datos para todos sus equipos, desde científicos de datos hasta gerentes de producto, lo que mejora la colaboración y la toma de decisiones basada en datos.
* **Mejora de la experiencia del usuario**: Los datos recopilados y analizados permiten a Spotify personalizar la experiencia del usuario, desde recomendaciones musicales hasta el desarrollo de nuevas funcionalidades en la plataforma.
* **Eficiencia operativa**: La infraestructura en la nube y las herramientas como Lexikon han reducido el tiempo necesario para procesar y acceder a datos relevantes, permitiendo a los equipos enfocarse en tareas de alto valor como la experimentación y el machine learning.
* **Cumplimiento normativo y gestión de costos**: Spotify ha implementado políticas claras de retención y acceso a los datos, lo que les permite cumplir con regulaciones y optimizar los costos de almacenamiento.

### **5. Conclusión**

* **Beneficios clave**: La plataforma de datos de Spotify, junto con herramientas como Lexikon, ha permitido a la empresa manejar de manera eficiente grandes volúmenes de datos, democratizar el acceso a insights y mejorar la toma de decisiones en toda la organización.
* **Lecciones aprendidas**: La iteración constante y la comprensión de las necesidades de los usuarios son claves para construir una plataforma de datos exitosa. Spotify ha logrado esto al adaptar su plataforma a las necesidades cambiantes de sus equipos y al fomentar una cultura de intercambio de conocimientos.
* **Futuro**: Spotify continuará mejorando su plataforma de datos y herramientas como Lexikon para seguir impulsando la innovación y ofreciendo una experiencia de usuario óptima.