Introducción

El siguiente es un proyecto de Ingeniería de Datos que busca determinar sus capacidades para desarrollar un proceso de ETL implementado sobre una arquitectura de BigData a partir de la especificación de un contexto de negocio. Se espera que Ud. desarrolle el proceso utilizando Spark.

Contexto de negocio
Una Fintech con operaciones en Latinoamérica, desea incrementar el engagement de los usuarios con su aplicación. Para ello, debe implementar una estrategia de negocio que le permita reducir la pérdida de usuarios registrados.

En un análisis de datos realizado por el equipo de User Experience (UX) se determinó que la mayor pérdida de usuarios registrados ocurre los primeros días de uso de la aplicación. Además, a partir del análisis de las respuestas de usuarios a las encuestas de UX, se formuló la hipótesis de que dicha pérdida de usuarios podría estar relacionada a la complejidad de uso de la aplicación, dada la gran cantidad de servicios que ofrece, lo que, probablemente se traduce en una experiencia con fuerte carga cognitiva.

En ese sentido, el equipo de negocio decidió hacer foco en los usuarios nuevos de la aplicación y propuso la implementación de una etapa de onboarding con 30 días de duración, en la que se guiará a los usuarios nuevos a crear sus productos bancarios, al mismo tiempo que se reducirá la carga cognitiva de la home page de la aplicación durante ese período.

Para medir la propuesta, los equipos de negocio y producto decidieron realizar un experimento de tipo A/B Testing con una serie de métricas que serán observadas durante un período de seis meses. Dichas métricas se detallan en la sección de experimentación.

Experimentación

Para llevar a cabo la medición de la etapa de onboarding, se decidió realizar durante seis meses un A/B Testing con los usuarios de Brasil (uno de los países con más usuarios de la aplicación). El grupo control estará formado por el 5% de los usuarios y el grupo tratamiento (aquellos que recibirán la nueva experiencia) por el 95%. Las métricas a observar durante la experimentación son:

1. **Drop**: Este indicador mide el porcentaje de usuarios que no vuelven a utilizar la aplicación después de ver la home page el mismo día de su registro (día en el que usuario crea la cuenta con todos sus datos y verificaciones).

2. **Activación**: Este indicador mide el porcentaje de usuarios en onboarding que realizan una primera acción transaccional.

3. **Hábito**: Este indicador mide el porcentaje de usuarios en onboarding que realizan al menos 5 transacciones. Aquí, es importante entender que existen dos segmentos de usuarios: individuals y sellers. Los individuals son usuarios que realizan sólo transacciones de pago, mientras que los sellers también realizan transacciones de cobro. El hábito para cada segmento se mide de la siguiente forma:
    a. Individuals: un usuario llega al hábito si realiza 5 transacciones en 5 días diferentes, durante el período de onboarding (fijado en 30 días).
    b. Sellers: un usuario llega al hábito si realiza 5 transacciones de cobro (sin importar los días), durante el período de onbarding (fijado en 30 días).

4. **Setup**: Este indicador mide el porcentaje de usuarios que realizaron al menos una acción de setup (por ejemplo. agregar una tarjeta de crédito a la billetera digital, activar la llave PIX, activar rendimientos en cuenta, etc.).

Datasets

Para realizar el proyecto se le proporciona un extracto histórico de los siguientes datasets:

- **lk_users**: contiene la información de los usuarios en onboarding para una fecha determinada.
- **bt_users_transactions**: contiene todas las transacciones de los usuarios de la lk_users realizadas durante un período determinado.
- **lk_onboarding**: contiene la información del usuario de onboarding que permite calcular las métricas de negocio para un período determinado.

A continuación una descripción de los campos más importantes presentes en los datasets:

1. **lk_users**:
    - **user_id**: identificador único de usuario. Tiene el prefijo “MLB” para referirse a usuarios de Brasil.
    - **rubro**: un identificador de rubro de negocio para usuarios del segmento seller.
2. **bt_users_transactions**:
    - **transaction_dt**: fecha de una transacción realizada por el usuario.
    - **type**: identificador de tipo de transacción. Para transacciones de pago se identifican del 1 al 7 y de cobro 8 y 9.
    - **segment**: el segmento al que pertenece el usuario: 1 para individuals y 2 para seller.

3. **lk_onboarding**:
    - **first_login_dt**: fecha histórica del primer inicio de sesión del usuario en la app.
    - **week_year**: día de la semana.
    - **user_id**: identificador unívoco anonimizado del usuario.
    - hábito: si es 1 indica presencia del atributo, caso contrario es 0.
    - **hábito_dt**: fecha en que realiza el evento de hábito.
    - activación: si es 1 indica presencia del atributo, caso contrario es 0.

- **activacion_dt**: fecha en que realiza el evento de activación.
- setup: si es 1 indica presencia del atributo, caso contrario es 0.
- **setup_dt**: fecha en que realiza el evento de setup.
- **return**: si es 1 indica que el usuario retornó posteriormente al first login.

Problema

Basado en el contexto de negocio y los datasets proporcionados, se le pide que utilice sus habilidades como Ingeniero de Datos para desarrollar un proceso de ETL usando Spark y Cassandra. La solución debe estar implementada sobre una arquitectura de Big Data y el pipeline de datos que se desarrolle debe proveer, al menos, una etapa de pre-procesamiento y una etapa de transformación (opcional una capa de visualización con un funnel que muestre la pérdida de usuarios por etapa de onboarding).

Para implementar la etapa de pre-procesamiento Ud. debe analizar con detalle los datos de cada uno de los datasets ya que pueden haber columnas innecesarias, valores perdidos o inconsistencia de datos (fuera de la lógica de negocio). La salida de esta etapa debe ser un conjunto de datos consistente acorde a la necesidad analítica de negocio.

Para la etapa de transformación Ud. debe poder preparar los datos para que puedan ser utilizados por científicos y analistas de datos. Se espera que el resultado de esta etapa permita realizar el análisis del desempeño de la etapa de onboarding para el grupo tratamiento del experimento.

Entregables

1. Una presentación donde se detalle la arquitectura, las tecnologías y todo el proceso de desarrollo: los recursos utilizados, propuesta de ingeniería para el contexto de negocio,las acciones realizadas en la etapa de pre-procesamiento y transformación, las decisiones tomadas y las conclusiones.
2. Un video explicativo basado en la presentación y en la solución realizada Spark y Cassandra. Debe mostrar el funcionamiento del pipeline y los resultados del ETL.

Evaluación

1. Diseño y desarrollo: Se evaluará el ingenio de la solución a partir del contexto de negocio, la arquitectura de la solución y la implementación en Spark y Cassandra (todos los datos deben estar en un Keyspace de Cassandra utilizando colecciones)..
2. Calidad: Correcto análisis de los datasets y ajuste de los datos de acuerdo a las reglas de negocio.
3. Profundidad: Decisiones tomadas y estrategias utilizadas.
4. Explicación del funnel de onboarding.
5. Storytelling: Se valorará el storytelling para presentarlo en el video.

Utiliza la información y los datos acorde a tu ingenio. Es una hoja en blanco y ¡vos podés proponer hasta donde te lleve tu creatividad!🚀