**Prueba técnica – Ingeniería de datos - Nequi**

Para la presente prueba técnica para ingeniero de datos, se establecieron algunos criterios para abordar de manera específica un problema desde la ingeniería de datos, los cuales se especifican a continuación junto a cómo se abordó cada uno:

1. **Alcance del proyecto y captura de datos:**

Se comenzó con la búsqueda y selección del dataset que cumpliera los criterios establecidos y se encontró un dataset con la simulación de la transaccionalidad de una entidad financiera, el cual incluía información de los clientes de cada una de las transacciones y las tarjetas correspondientes (<https://www.kaggle.com/datasets/ealtman2019/credit-card-transactions>).

Posteriormente, se exploraron los datos y a partir de allí se definieron los casos de uso final, los cuales son los siguientes:

* Se entenderá quiénes son los clientes que tiene la entidad financiera, y cuál es su comportamiento transaccional, para ver si se alinean con los objetivos organizacionales y encontrar puntos de mejora. Para ello se harán algunas gráficas que nos detallen las posibles correlaciones entre variables, distribuciones y datos atípicos en casos específicos.
* Se obtendrá una fuente de datos de la verdad, ya que se hará una limpieza y una validación de los datos de manera tal que estén listos para el consumo de diferentes áreas dentro de la entidad financiera.
* Obtener una alta cantidad de características (feature engineering), para proporcionar suficientes datos para dar mayor fuerza al entrenamiento de modelos de ML.

1. **Explorar y evaluar los datos a partir del EDA:**

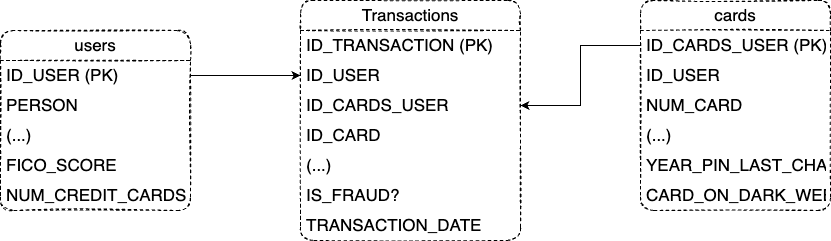
Para la exploración y la evaluación de los datos obtenidos, se realiza un análisis de datos exploratorio (EDA), en el cual se entiende de manera general cómo es el comportamiento y la naturaleza de la mayoría de los datos obtenidos para así:

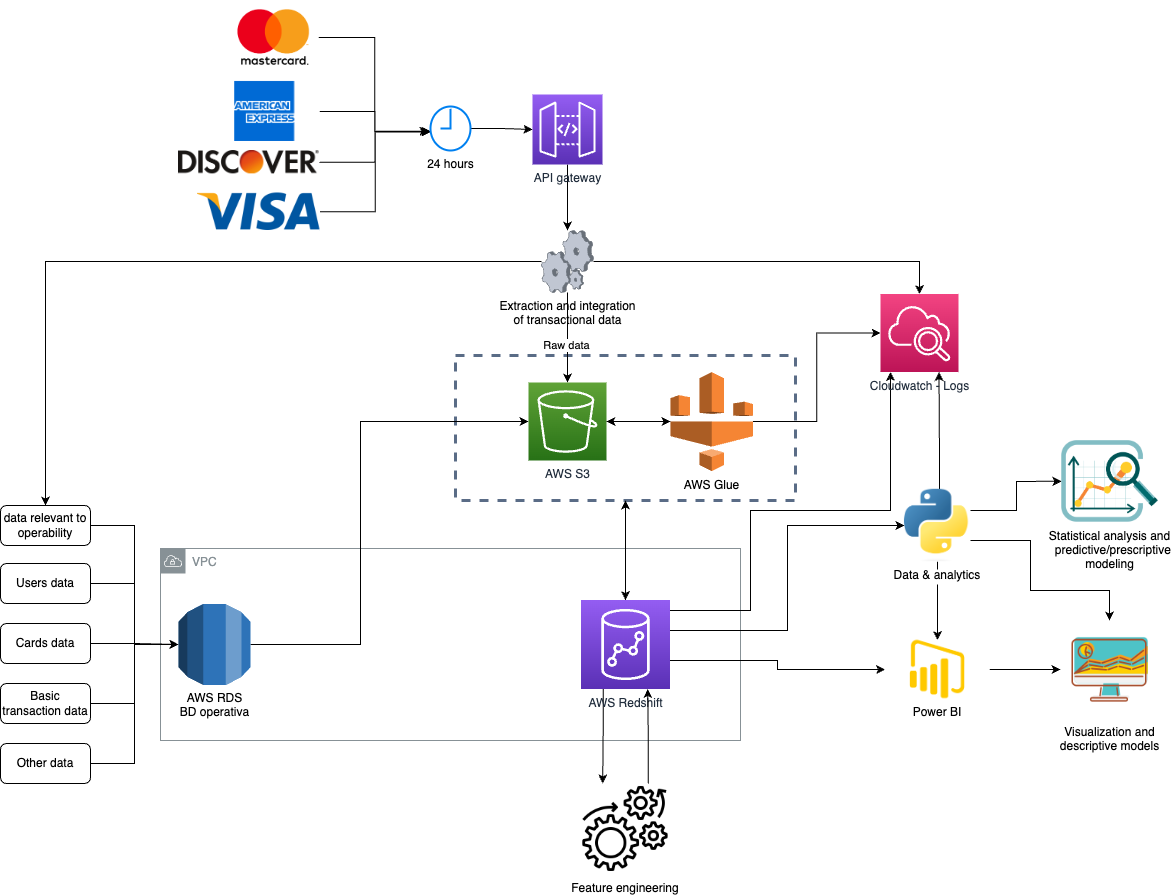
* Detectar y corregir posibles errores de formatos de atributos
* Detectar y administrar los valores vacíos de cada atributo
* Detectar y administrar los valores duplicados
* Detectar y agregar nuevas columnas que den nuevos insumos para el entendimiento de los datos
* Detectar datos atípicos

1. **Definir el modelo de los datos:**

* Trazar el modelo de datos conceptual y explicar por qué se eligió ese modelo.

R:// Con los datos que se recolectaron, se puede generar el siguiente modelo, que se asemeja a un modelo estrella, el cual nos proporciona todas las características necesarias para cumplir los objetivos establecidos.



* Diseñar la arquitectura y los recursos utilizados.   
  R:// Para el diseño de la arquitectura se aprovechó las herramientas cloud, específicamente AWS, el cual es un proveedor cloud robusto y que provee gran diversidad de herramientas para la gestión de recursos. A continuación se muestra la arquitectura propuesta y posteriormente la explicación de cada uno de los componentes que se usaron:  
    
  

Para la anterior arquitectura se usan los siguientes componentes:

* **AWS RDS**: Para el alojamiento y la administración de la base de datos operativa.
* **AWS API** Gateway: Para disparar el servicio de extracción e integración de los datos transaccionales entregados por las franquicias de tarjetas de pago.
* **AWS S3**: Para almacenamiento de datos en general
* **AWS** Glue: Para la detección, preparación, migración e integración de datos provenientes de varios orígenes para el análisis
* **AWS Redshift**: Para disponibilizar datos de diferentes fuentes de una manera eficaz y sin sobrecargar la base de datos operativa.
* **Feature engineering**: Para la creación de características que aporten valor a los modelos
* **Power BI**: Para creación de dashboards
* **AWS Cloudwatch**: Para el almacenamiento y administración de logs

Para el caso específico se tiene el supuesto de requerir una actualización diaria, sin embargo, esto dependerá de la inmediatez que requieran tomar acciones y el uso en general que se les darán a los datos, como por ejemplo:

* + En caso tal de querer detectar movimientos anómalos por algún tipo de fraude, la actualización debe ser con una corta periodicidad, llegando a ser a tiempo real (Aunque para ello, las plataformas de las franquicias proveedoras de las tarjetas proporcionan esta información a tiempo real) o con una actualización como mínimo diariamente.
  + En caso tal de requerir tener datos para objetivos comerciales del día a día, se podría actualizar diariamente los datos.
  + En caso tal de que sea para la definición de estrategias, se podría establecer una actualización semestral, trimestral o mensual.

1. **Ejecutar el ETL:**

Para la creación del modelo de datos, se realizó un ETL básico, el cual cumple con los siguientes criterios:

* Crear las tuberías de datos y el modelo de datos
* Ejecutar controles de calidad de los datos para asegurar que la tubería funcionó como se esperaba
* Control de calidad en los datos con la integridad en la base de datos relacional (por ejemplo, clave única, tipo de datos, etc.)
* Pruebas de unidad para los “Script” para asegurar que están haciendo lo correcto.
* Comprobaciones de fuente/conteo para asegurar la integridad de los datos.
* Incluir un diccionario de datos
* Criterio de reproducibilidad

1. Completar la redacción del proyecto**:**

* ¿Cuál es el objetivo del proyecto?  
  R:// El objetivo inicial del proyecto es identificar oportunidades para aumentar la transaccionalidad de la entidad financiera, lo cual se puede hacer revisando cómo se están segmentando las transacciones basado en algunas características.
* ¿Qué preguntas quieres hacer?   
  R:// Las preguntas que se pueden resolver con el trabajo expuesto son sobre cómo se están segmentando nuestros clientes,
* ¿Por qué eligió el modelo que eligió?   
  R:// Se eligió este modelo ya que era el más adecuado basado en los datos que obtuve y el tiempo estimado para la prueba.
* Incluya una descripción de cómo abordaría el problema de manera diferente en los siguientes escenarios:
  + Si los datos se incrementaran en 100x.  
    R:// Para la extracción de los datos se podría implementarían técnicas de mapReduce y revisar si los costos de Redshift siguen siendo atractivos o implementar Athena para la consulta de estos datos.  
    Para la analítica en general, se podría utilizar procesamiento en paralelo y utilizar AWS sagemaker para procesar de manera más rápida los datos y los modelos.
  + Si las tuberías se ejecutaran diariamente en una ventana de tiempo especifica.  
    R:// Establecería un cron para la ejecución diaria y escogería la hora que menos tráfico tenga, para no sobrecargar la base de datos operativa.
  + Si la base de datos necesitara ser accedido por más de 100 usuarios funcionales:  
    R:// Redshift estaría en capacidad de entregar estos datos, sin embargo, lo ideal sería particionar de acuerdo con las necesidades de cada usuario funcional y así entregar sólo la información necesaria, dando cumplimiento al gobierno de datos.
  + Si se requiere hacer analítica en tiempo real, ¿cuáles componentes cambiarias a su arquitectura propuesta?   
    R:// Para ello comenzaría a hacer toda la captura, procesamiento y almacenamiento con herramientas cloud específicas para ello, como lo es Amazon Kinesis Data Streams y Amazon Kinesis Data Analytics en conjunto con servicios serverless como AWS Lambda para ejecutar de manera asíncrona algunos microservicios.