Prueba técnica - Ingeniero de datos

Fecha: 09 diciembre 2024

Nombre: Manuela Suescun Serna

Cédula:1193219656

La siguiente prueba técnica es abierta, los entregables corresponden al código fuente, un README con el detalle de implementación y justificación de diseño y un informe cuyo contenido se detalla más adelante. Esto se debe entregar en un repositorio git a elección.

Fecha límite de entrega: Lunes, 09 de diciembre de 2024, hora: 12:00 medio día

¡Muchos éxitos!

Descripción:

Se ha suministrado un archivo llamado ***Films2.xlsx*** donde encontrará el MER (Modelo Entidad - Relación) compuesto por 5 tablas, en cada pestaña del archivo se encuentra la información de cada una de las tablas.

Tener en cuenta el archivo *Films2* para formular preguntas relevantes que puedas responder utilizando las tablas del archivo. Adicionalmente, tener en cuenta los siguientes requisitos:

1. Implementar una aplicación que permita gestionar y ejecutar ETLs utilizando Python como lenguaje de programación. Esta aplicación debe ser escalable e implementar módulos de observabilidad. El paradigma de programación es a elección, pero es preferible POO.
2. Se deben implementar los ETL en esta aplicación usando un framework de procesamiento de datos (preferible Spark).
3. (Opcional) Desplegar la prueba en herramientas de la nube (GCP, AWS, Azure).
4. Realizar informe de presentación de resultados (max 3 páginas):
    1. Explicación de la arquitectura de datos y arquetipo de la aplicación.
    2. Análisis exploratorio de datos.
    3. Formular 5 posibles preguntas de negocio a las cuales los datos procesados puedan dar respuesta. Responde las preguntas.
    4. Conclusiones.

Nota: Es importante verificar la información, con el propósito de comprobar su consistencia. Dado el caso de encontrar irregularidades, por favor realizar un proceso de limpieza.

1. Arquitectura de Datos y Arquetipo de Aplicación

Arquitectura de Datos: Basándonos en las tablas proporcionadas, podemos ver que siguen un modelo relacional. Cada tabla representa una entidad específica:

- **film:** Información detallada sobre las películas, como duración, clasificación, características especiales.
- **inventory:** Registra las copias disponibles de cada película en cada tienda.
- **rental:** Detalla los alquileres realizados, incluyendo la fecha de alquiler y devolución.
- **customer:** Contiene información sobre los clientes, como nombre, dirección y estado de la cuenta.
- **store:** Informa sobre las tiendas, incluyendo el empleado encargado.

Las relaciones entre estas tablas son las siguientes:

- **film** y **inventory:** Una película puede tener múltiples copias en el inventario.
- **rental** y **inventory:** Cada alquiler se asocia con una copia específica del inventario.
- **rental** y **customer:** Cada alquiler está vinculado a un cliente.
- **store** y **inventory:** Las copias de las películas están ubicadas en una tienda específica.
- **store** y **customer:** Los clientes están asociados a una tienda específica.

Arquetipo de Aplicación: Este conjunto de datos parece ser parte de un sistema para gestionar videoclubes o quizás una plataforma de streaming. La aplicación que lo respalda funciona como un sistema de registro transaccional que lleva el control de las operaciones diarias relacionadas con el alquiler de películas.

2. Análisis Exploratorio de Datos (EDA)

Observaciones Iniciales:

- **Completitud de los datos:** Las tablas parecen tener una estructura consistente y sin valores nulos evidentes en los datos proporcionados.
- **Tipos de datos:** Los datos son una mezcla de numéricos (por ejemplo, rental_duration, rental_rate) y categóricos (por ejemplo, rating, special_features).
- **Relaciones:** Las relaciones entre las tablas son claras y permiten realizar análisis detallados sobre los patrones de alquiler.

Preguntas de Investigación Iniciales:

- ¿Cuáles son las películas más populares (basadas en el número de alquileres)?
- ¿Cuál es la duración promedio de alquiler?
- ¿Cuál es el género de película más alquilado?
- ¿Existe una relación entre la calificación de una película y su popularidad?
- ¿Cuáles son los clientes más activos?

3. Preguntas de Negocio y Respuestas

Pregunta 1: ¿Cuál es la película más alquilada y cuál es la menos alquilada?

- **Respuesta:** Para responder a esta pregunta, necesitaríamos unir las tablas rental y film por el campo film_id. Luego, agrupar por film_title y contar el número de alquileres. La película con el mayor número de alquileres sería la más popular.

Pregunta 2: ¿Cuál es el género de película más rentable?

- **Respuesta:** Suponiendo que tenemos un campo genre en la tabla film, podríamos calcular el ingreso total por género multiplicando la tarifa de alquiler por el número de alquileres por género.

Pregunta 3: ¿Cuál es la duración promedio de alquiler de las películas clasificadas como 'G'?

- **Respuesta:** Filtraríamos los datos de la tabla rental por películas clasificadas como 'G' y calcularíamos la duración promedio de los alquileres.

Pregunta 4: ¿Existe una correlación entre la duración de una película y su tarifa de alquiler?

- **Respuesta:** Calcularíamos el coeficiente de correlación de Pearson entre las columnas length y rental_rate de la tabla film para determinar si existe una relación lineal entre ambas variables.

Pregunta 5: ¿Cuál es el cliente que ha generado más ingresos para el videoclub?

- **Respuesta:** Uniríamos las tablas rental y customer por customer_id. Luego, calcularíamos el ingreso total generado por cada cliente y encontraríamos al cliente con el ingreso más alto.

Conclusiones

El análisis de este conjunto de datos permite obtener insights interesantes:

- **Popularidad de las películas**: Identificar cuáles se alquilan más o menos.
- **Preferencias de los clientes**: Detectar los géneros y clasificaciones más demandados.
- **Desempeño de las tiendas**: Comparar alquileres e ingresos entre tiendas.
- **Estrategias de precios**: Evaluar si el precio influye en la cantidad de alquileres.

Siguientes pasos con ETL:

1. **Limpieza de datos**: Corregir errores o inconsistencias.
2. **Análisis avanzado**: Usar técnicas estadísticas y visualizaciones para profundizar en los datos.
3. **Predicciones**: Crear modelos para anticipar la demanda de películas o el comportamiento de los clientes.

Recomendaciones:

- **Agregar más datos**: Incluir información como edad o género de los clientes, así como datos sobre promociones.
- **Visualizaciones**: Crear gráficos interactivos con herramientas como Power BI o Tableau.
- **Análisis temporal**: Explorar cómo evolucionan los alquileres con el tiempo para encontrar patrones o tendencias.