Data Arch & Governance

Prueba de técnica
Mercado Libre

Contexto

Estimado candidato, esta prueba tiene como objetivo evaluar sus capacidades técnicas en el ámbito de desarrollo tecnológico en el área de datos, por lo cual tiene libre elección respecto a las herramientas tecnológicas y lenguaje de programación con los cuales va a desarrollar, cabe resaltar que es muy importante que nos comparta en que tecnologías resolvió los problemas aquí planteados acompañado a su vez de un diseño y su respectiva documentación.

Objetivo

Esta prueba consta de tres puntos de desarrollo, cada uno con fines y objetivos diferentes con los cuales buscamos poner a prueba sus habilidades técnicas en diseño y programación.

Tiempo: El tiempo estimado para el desarrollo de prueba es de 7 días (1 Semana) después de la recepción de la presente prueba.

Entrega: Se debe entregar código, repositorios, diseños, documentos y cualquier otro entregable que se considere necesario para el entendimiento de la solución de la misma.

Canal de comunicación: Cualquier duda no dude en contactarnos a los siguientes correos: data-arch-governance@mercadolibre.com, edison.neira@mercadolibre.com.co

Punto 1

A partir de de los data set compartidos (compras.csv, salarios.csv) responde las siguientes preguntas:

1. De acuerdo al set de datos SALARIOS ¿Cuántos cargos estaban ocupados solamente por una persona en 2011?
2. De acuerdo al set de datos SALARIOS ¿Cuánta gente tiene la palabra 'MANAGER' en su cargo?
3. De acuerdo al set de datos SALARIOS ¿Cuál es el nombre de la persona que menos gana (incluyendo beneficios - TotalPayBenefits)?
4. De acuerdo al set de datos SALARIOS¿Cuál es el salario base (BasePay) promedio de todos los empleados para el año (2012)?
5. De acuerdo al set de datos SALARIOS ¿Cuál fue la suma total pagada con beneficios por los dos trabajos más populares?.
6. De acuerdo al set de datos COMPRAS ¿Cuáles son los 5 proveedores de correo electrónico más comunes, con cuantos usuarios está asociado cada uno? (hotmail.com,gmai.com, etc)
7. De acuerdo al set de datos COMPRAS ¿Cuántas personas tienen una tarjeta de crédito que expira en 2025?
8. De acuerdo al set de datos COMPRAS ¿Cuántas personas tienen tarjetas Mastercard e hicieron una compra por más de $20?
9. De acuerdo al set de datos COMPRAS ¿Alguien hizo una compra desde Lot: "90 WT", ¿cuál fue el precio de compra de esta transacción?
10. De acuerdo al set de datos COMPRAS ¿ Cuánto suma el total de precio de compras para las dos compañías menos populares?, ¿Cuáles son esas dos compañías?

Nota: Es importante que nos compartas el código y/o consultas usadas para responder estas preguntas junto con el diseño y tecnologías usadas para responderlas.

Punto 2

En Meli tenemos el desafío de procesar tablas de clientes por transacción, dicha información reposa en dos rutas diferentes en formato csv particionado (Ver dataset punto 2), a partir de estos datos la idea es generar un insight (Tabla, Data Frame, etc) de **clientes únicos** existentes con las siguientes columnas: **TIPO DE IDENTIFICACION, NUMERO DE IDENTIFICACION, TIPO+NUMERO DE IDENTIFICACION (Columna nueva), HASH(Columna nueva)**. El objetivo de este insight es tener un set de datos de solo clientes únicos, en este caso se debe ejecutar las operaciones que considere necesario para cumplir este objetivo(Crear las columnas necesarias, por ejemplo la hash). Es importante resaltar que la columna **hash** saldrá del campo concatenado [**TIPO+NUMERO DE IDENTIFICACION**] y se debe usar algún algoritmo de hashing de su preferencia (SHA,AES, MD5, etc).Por otra parte se debe construir un insight (Tabla, Data Frame, etc) de **transacciones** donde se elimine las columnas sensibles y únicamente cuente con las columnas **VALOR_TX** y **HASH**, está columna hash es la misma resultante de la concatenación que se dio en el insight de clientes pues esta es la columna con la que se permite realizar relación entre el insight de clientes y transacciones. Para finalizar la prueba se debe realizar una consulta que retorne el top **10** de los clientes con más transacciones realizadas, cabe recordar que la búsqueda se debe realizar por la clave hash en el insight de transacciones

Nota: El entregable de este punto es el diseño técnico de la solución, la solución implementada y tecnologías usadas (Preferiblemente en cloud).

Punto 3

En este punto contamos con dos archivos heterogéneos en formato (bookCatalog.xml, datos_compras_v2.csv) y un api rest (https://restcountries.com/) en cual se describen libros, compradores de estos libros y países donde viven los compradores; en este punto debe obtener insights importantes para la empresa, por tal razón es importante que analice la información suministrada y proponga qué preguntas de negocio se pueden plantear y desarrollar su respectiva respuesta (ETL), además de compartir las conclusiones obtenidas a partir de la información suministrada y que proceso de análisis, enriquecimiento, transformación y visualización de datos fue ejecutado para la resolución a partir de los data sets suministrados.

El entregable de este punto es el desarrollo de la solución en donde se evidencia que preguntas de negocio se plantearon y resolvieron, importante entregar documentación de tecnología y plataformas usadas para la solución de este punto.
Bonus: Mockups de posibles visualizaciones que se pueden realizar a partir de este set datos.

No duden en escribirnos por si tienen alguna duda sobre la prueba y con gusto les daremos respuesta.

¡¡¡Muchos éxitos!!!!