Diseño de un proyecto de ciencia de datos

1. Arquitecturas típicas de proyectos de datos masivos.

Campaña para venta de seguros

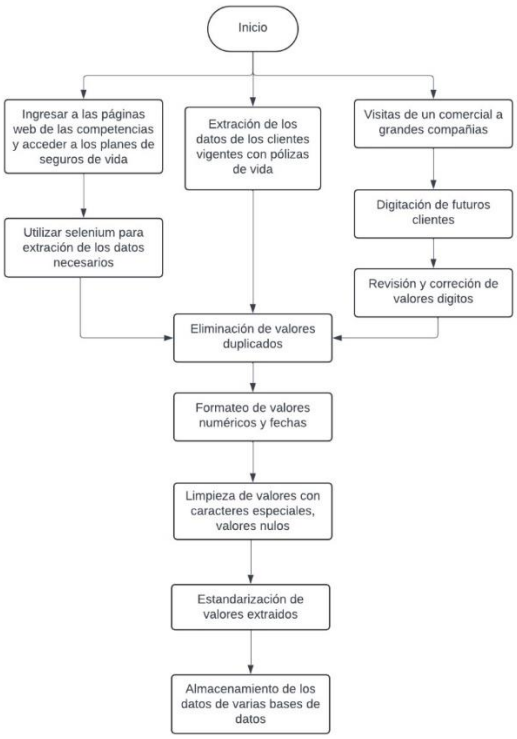
Fuentes heterogéneas:

1) Histórico de los clientes de la compañía de seguro: Base de datos que la compañía tiene para analizarlos.

2) Benchmarking de la competencia: Web scraping de los planes y tipo de seguros de la competencia para utilizarlo de comparativa para referencia a nuestra compañía.

3) Datos obtenidos por vendedores: Conversaciones con grandes compañías, las cuales puedan requerir de un seguro de vida para sus empleados para obtener un listado de potenciales clientes.

Extracción, transformación y carga (ETL):



Almacenamiento:

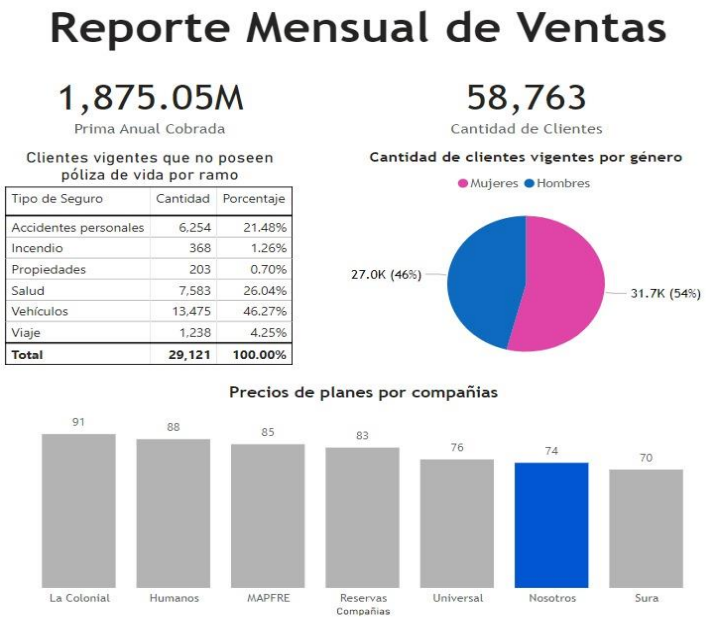
No SQL está descartado porque la información es estructurada, el caso de los Data Lakes son servicios en la nube que regularmente las compañías tienen sus propios almacenes de datos porque así pueden tener más seguridad y manejo de su información

por tanto el utilizaremos el Dataware house ya que es centralizado, lo cual permite tener todos los datos en un único lugar, facilitando el acceso, la gestión, permite análisis de grandes volúmenes de datos con herramientas que se pueden integrar a este mismo, es seguro, se pueden configurar controles de acceso, políticas se seguridad y permite su actualización continua.

Tratamiento de los datos:

1) Eliminación de valores nulos
2) Eliminar duplicados
3) Formateo de dato numérico y fecha
4) Estandarizar todas las columnas
5) Conexionar DataWare House
6) Hacer join en caso de que se necesite
7) Elaboración de Eda

Visualización:



2. Perfil del científico de datos.

Ciencias de la computación:

Nuestro equipo está compuesto por un analista líder Python, SQL, R, Power Bi, Tableau, Power Query, liderazgo, conocimiento de técnicas y estadística , solución de conflicto y Kanban y los analistas de datos Python, SQL, R, Power Bi, Tableau, MatPlotLib, Seaborl, Excel, manejo de base de datos relacionales (SQL), conocimiento de herramienta de machine learning Scikit-Learn , Keras y SAS.

Matemáticas.:

Análisis de frecuencia acumulada, de regresión, de varianza, correlación, la media, desviación estándar, determinar el tamaño de nuestra muestra, prueba de hipótesis, entre otras.

Comunicación:

En el dashboard se encuentra la información resumida y se pudiese elaborar una presentación para resumir los aspectos más importantes de la data, analizarla y visualizar que tipo de herramienta pudiésemos usar para eso.

Negocios:

Captar una mayor de clientes para así poder incrementar las ventas y por ende el mismo negocio.

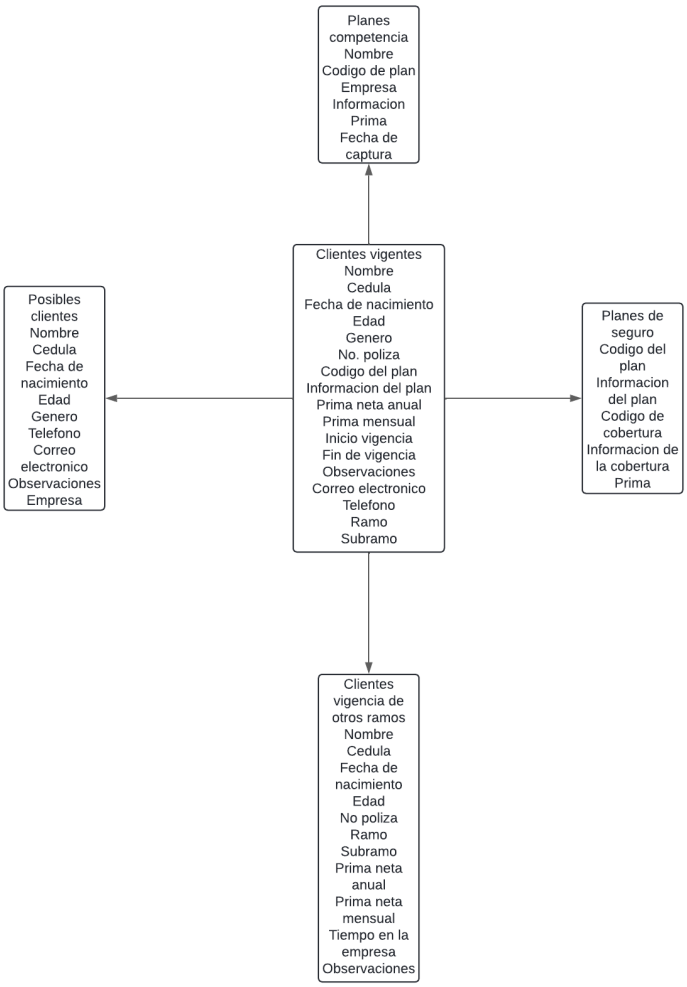
Obtener un precio competitivo en el mercado.

Obtener planes para entierros, en conjunto con el seguro de vida, permitiendo que el cliente pueda tener varios beneficiarios.

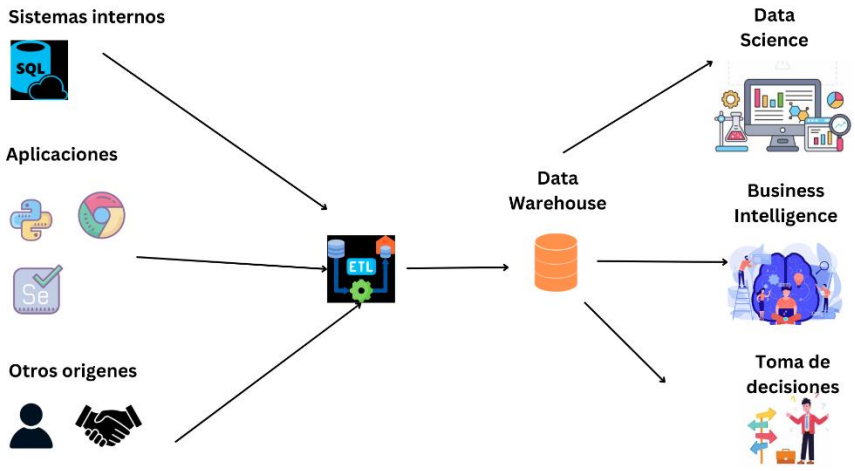
Obtener planes asociados incluyendo un seguro de salud junto con el seguro de vida

3. Estrategias en almacenamiento masivo.

Data Mart:



Data Warehouse:



Data Lake

Permiten el acceso inmediato a todos los datos recopilados

Al no ser información estructurada no es necesario analizarla ya que se realizará posteriormente bajo demanda

Sirve para procesos de machine learning, análisis predictivo y creación de perfiles

Se puede almacenar grandes volúmenes de datos de manera escalable

Nuevas tendencias en almacenamiento masivo:

Ingesta de datos no estructurados:

Se pueden usar herramientas y conectores, por ejemplo, Azure Data Factory para la creación de flujos de trabajos a gran escala, con el que se pueda mover, transformar y procesar datos de diversas fuentes como correos, documentos, imágenes, etc.

Extracción de datos con OCR (Reconocimiento óptico de caracteres) para convertir imágenes y documentos escaneados en texto estructurado como los reclamos de clientes, contratos de las pólizas, etc.

Usar herramientas como Azure Data Catalogó para organizar y gestionar los datos no estructurados y procesarlos para su futuro uso.

Crear Pipelines de ingesta que automaticen la captura, procesamiento y almacenamiento de los datos, con los que se puedan almacenar de manera más rápida en el Data Lake.

Uso de Hive para consultas SQL de los datos no estructurados almacenados en clústers de Hadoop.

4. Estrategias de aplicación de la ciencia de datos y datos masivos.

Inteligencia de negocio:

En base a los datos obtenidos de la data histórica vamos a poder hacer varios análisis y en base a estos análisis tomar las decisiones pertinentes para el negocio y que esto pueda influir directamente para obtener una mayor cantidad de clientes y un rango de competencia por encima del promedio.

Analítica de negocio:

En relación con los datos obtenidos del Webscraping pudimos observar que los precios de los planes de la competencia son mas caros que el nuevo plan que vamos a lanzar y se pudiese elaborar una campana de venta para conocer el plan en conjunto a los beneficios que podemos ofrecer como servicios funerarios y para clientes nuevos aproximadamente un 10% de descuento y de esta manera poder captar una mayor cantidad de clientes.

Minería de datos:

Se pudiesen usar arboles de decisión para la clasificación de clientes en base a una serie de parámetros, clustering para agrupar clientes en diferentes tipos según ciertas características que puedan tener y con redes neuronales se pudiese saber que tipo de estrategias aplicar a cada grupo de clientes.

Aprendizaje automático:

Utilizamos Python con librerías de aprendizaje automático para analizar y comparar el rendimiento de las estrategias, como técnicas de regresión lineal o regresión logística con los que se puedan predecir las ventas en función de las variables independientes. Además de algoritmos de clasificación como árboles de decisión o bosque aleatorio con los que se puedan segmentar los clientes que puedan comprar un seguro y hacer agrupaciones de clientes por segmentos con técnicas como K-Means.

Inteligencia artificial:

Aquí también utilizaríamos Python para implementar técnicas de Inteligencia Artificial con Keras y TensorFlow con redes neuronales que nos permitan analizar grandes conjuntos de datos e identificar tendencias y patrones en nuestros datos, hacer cálculos de manera intensiva y eficiente.