
# Propuesta de trabajo

## Empresa: Olist

Castro P. - Fernandez G. - Pierotti L. - Tachella F.

# Índice

# Introducción

Nuestra empresa ha diseñado un plan de acción para conectar a las PYMEs con mercados más amplios, mejorando la experiencia del usuario. Este esquema incluye una investigación exhaustiva del mercado de comercio electrónico en Brasil con el objetivo de desarrollar estrategias que faciliten la inteligencia del negocio y encontrar soluciones innovadoras para ayudar a los usuarios a vender sus productos a una base de clientes más amplia. Utilizaremos información obtenida a través de una fuente de acceso libre conocida como "Olist" para llevar a cabo estas operaciones.

# Conformación del equipo: Roles y funciones

Para maximizar la eficiencia de este proyecto, las tareas serán distribuidas entre los miembros del equipo en función de sus habilidades y preferencias individuales. Se permitirá una mayor flexibilidad en la asignación de tareas, permitiendo a los miembros del equipo trabajar en diferentes áreas según la complejidad del proceso. Además, los avances serán supervisados por todos los miembros del equipo para garantizar el progreso del proyecto.

Data Engineer: Lautaro y Guillermo

Data Analytics: Franco y Pablo

Machine Learning: Guillermo

# Metodología de trabajo

El proyecto se gestionará utilizando la metodología Scrum. Se llevará a cabo una reunión diaria de equipo sin límite de tiempo para discutir posibles cambios en el proyecto y seguir el progreso de los objetivos establecidos. Además, al final de cada semana se realizará una revisión global de los avances del proyecto. También se programaron reuniones regulares con el cliente para informarle del progreso del trabajo y recibir su retroalimentación. El equipo tendrá un cronograma de tareas, pero estará abierto a adaptaciones en función de las necesidades y las ideas que surjan a lo largo del proceso, tanto del equipo como del cliente.

# Objetivos

### Objetivo principal:

- Conectar a pequeñas empresas (PYMEs) con mercados más grandes y mejorar la experiencia del usuario.

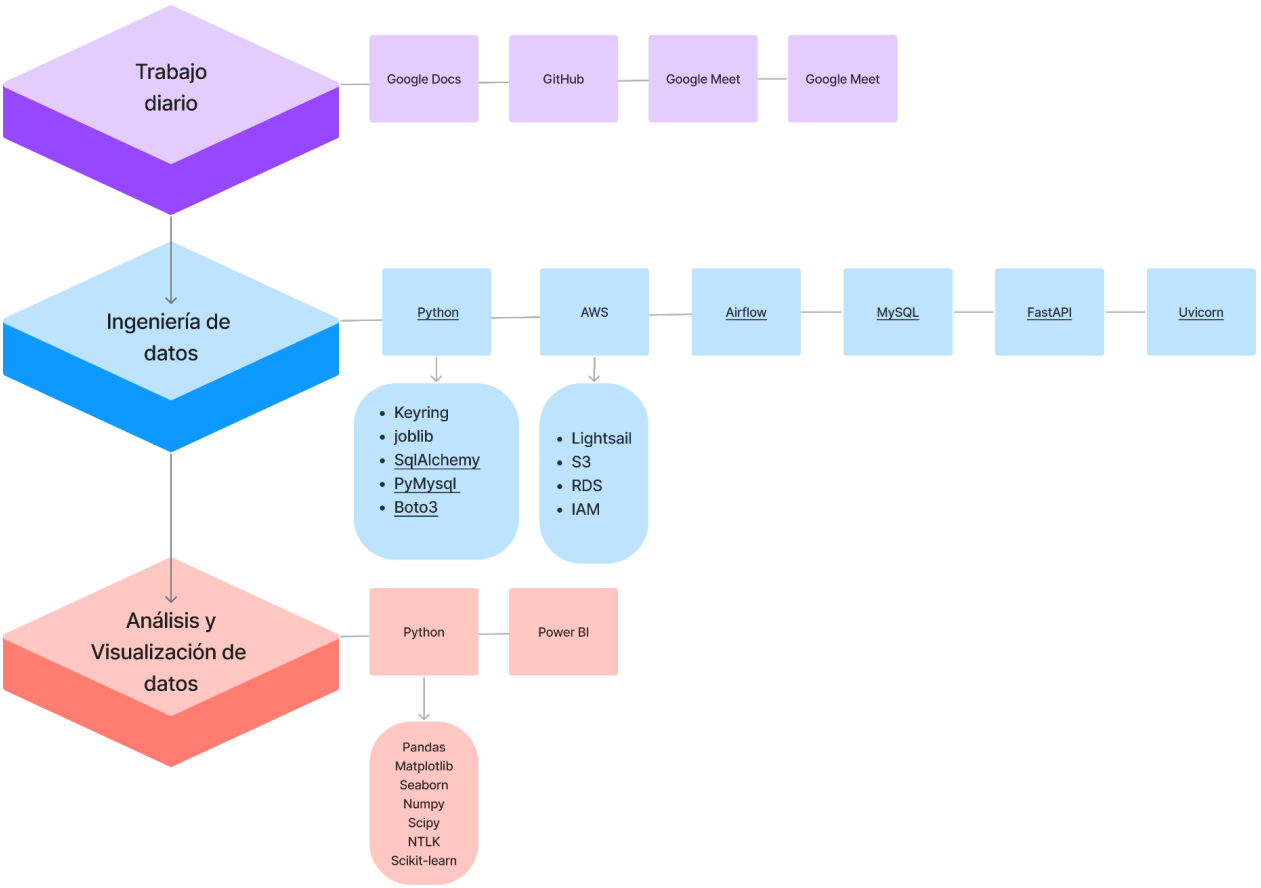
### Objetivos específicos:

- Construir un Data Lake y Data Warehouse eficiente y escalable.
- Analizar a vendedores y su performance.
- Extraer KPIs que engloban el performance de cada vendedor.
- Ofrecer recomendaciones a la empresa en base a los datos disponibles.
- Armar un reporte para visualización escalable y en línea.

# Alcances del proyecto

En este proyecto se busca proporcionar al cliente distintas herramientas para trabajar con sus datos y a su vez también un análisis de los datos disponibles, con los cuales se buscará ofrecer recomendaciones al cliente para la mejora de sus servicios.

# Stack tecnológico propuesto:



# Cronograma de trabajo

## SEMANA 1

Durante la primera semana del proyecto nos enfocaremos en establecer un plan detallado para llevar a cabo el trabajo y en recopilar y analizar los datos necesarios para llevar a cabo el proyecto. Es importante tener una buena comprensión de los datos antes de comenzar a trabajar en el desarrollo del proyecto.

### Lunes y Martes:

- Análisis y entendimiento del proyecto
- Definición de objetivos y alcance

### Miércoles y Jueves:

- Análisis exploratorio de los datos (EDA)

### Viernes:

- Reunión con el product owner y propuesta de proyecto

## SEMANA 2

### Sábado y Domingo:

- Definir Infraestructura y elección del servicio en la nube

### Lunes, Martes y Miércoles:

- Creación del bucket de S3 y la base de datos en amazon RDS
- Armado de airflow en la máquina virtual de amazon lightsail

### Jueves:

- Testeo de la infraestructura en la nube y últimos cambios

Viernes:

- Presentación de la infraestructura funcionando al product owner

## SEMANA 3

Sábado a Miércoles:

- Construcción de un dashboard preliminar
- Armado de modelo de machine learning

Jueves:

- Organizado de storytelling

Viernes:

- Presentación del dashboard preliminar al product owner

## SEMANA 4

Sábado y domingo:

- Finalización del dashboard
- Análisis y búsqueda de mejoras para ofrecer recomendaciones
- Armado de powerpoint para presentacion final

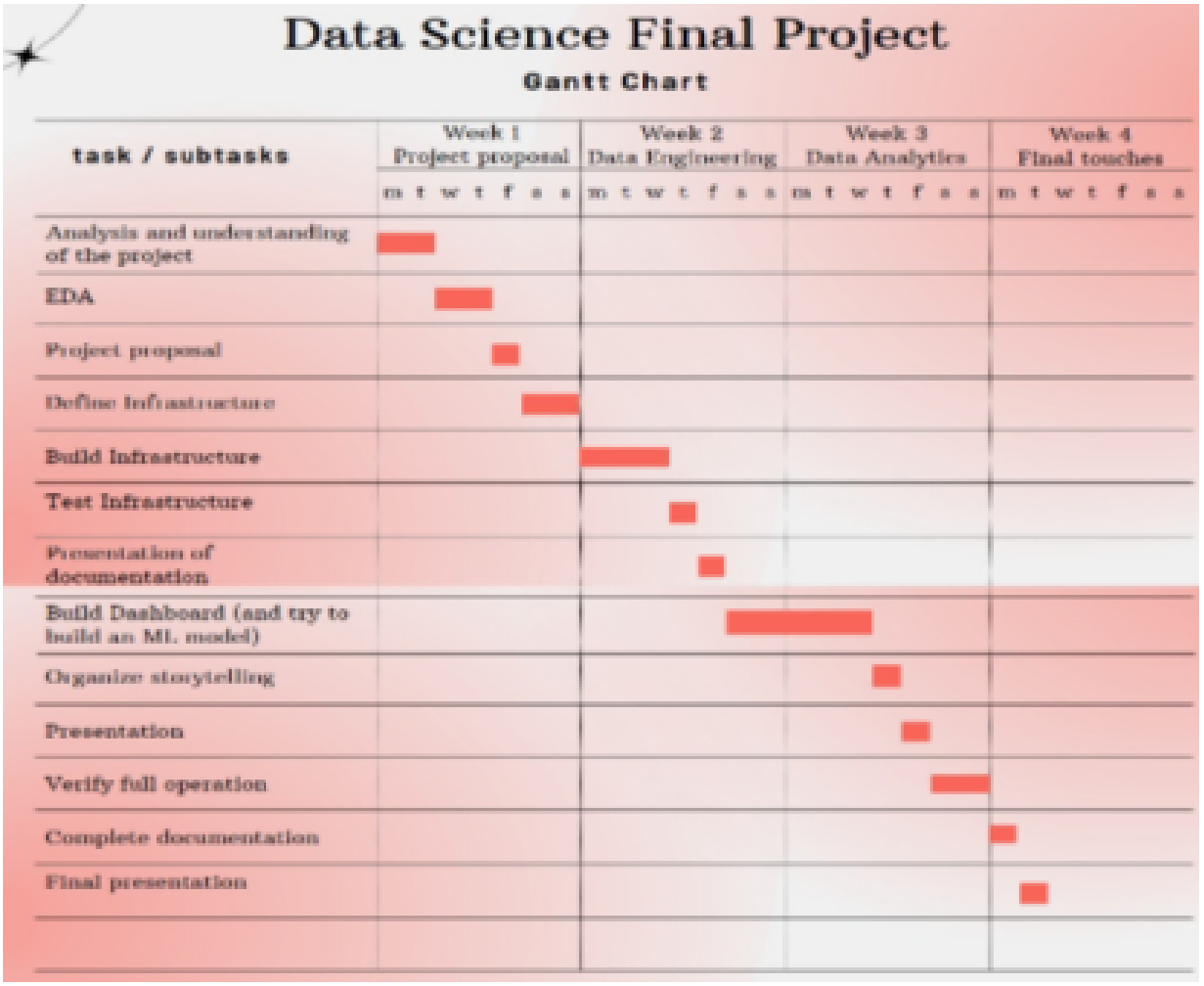
Lunes:

- Reunión final para organizar storytelling y finalizar detalles

Martes:

- Presentación del proyecto

# Estimación de esfuerzos



# Entregables

- Todo el código organizado en un repositorio en GitHub
- Base de Datos escalable
- ETL automatizado
    - Carga auto-incremental configurada
    - Tablas de auditoría
    - Tablas nuevas creadas para el análisis
- Análisis de vendedores y su desempeño
- Recomendaciones de negocio
- KPIs extraídos
- Dashboard de Power BI
- Modelo de ML sobre tiempos de entrega y reviews
- API para consultas de performance de vendedores

# Desarrollo del proyecto

## Diccionarios de datos

| Dataset | Closed_deals | | |
|---|---|---|---|
| Columna | Detalle | Tipo de dato | Ejemplo |
| mql_id | ID cliente potencial | Alfanumerico | a0604c9d9ef23fbf7cb7be5091201041 |
| seller_id | ID del vendedor | Alfanumerico | b7140ce94c4514bf136a2c3f98e0476c |
| sdr_id | informacion incompleta | Alfanumerico | b90f87164b5f8c2cfa5c8572834dbe3f |
| sr_id | informacion incompleta | Alfanumerico | d3d1e91a157ea7f90548eef82f1955e3 |
| won_date | informacion incompleta | Fecha y hora | 2018-07-31 20:01:32 |
| business_segment | Segmento del negocio | Texto | audio_video_electronics |
| lead_type | Tipo de cliente | Texto | online_small |
| lead_behaviour_profile | informacion incompleta | Texto | cat |
| business_type | Tipo de negocio | Texto | reseller |
| declared_monthly_revenue | informacion incompleta | Numerico decimal | 0.0 |
| | | | |
| Dataset | Customers | | |
| Columna | Detalle | Tipo de dato | Ejemplo |
| customer_id | ID cliente | Alfanumerico | bc3c9c45fe3fd83f49adbcbf50daa3da |
| customer_unique_id | ID unico del cliente | Alfanumerico | e72bbc364013bd2f23ace1b4e3c43be6 |
| customer_zip_code_prefix | Codigo postal del cliente | Numerico entero | 89252 |
| customer_city | Ciudad del cliente | Texto | jaragua do sul |
| customer_state | Estado del cliente | Texto | SC |
| | | | |
| Dataset | Geolocation | | |
| Columna | Detalle | Tipo de dato | Ejemplo |
| geolocation_zip_code_prefix | Codigo postal | Numerico entero | 31035 |
| geolocation_lat | Latitud | Numerico decimal | -19.898.344 |
| geolocation_lng | Longitud | Numerico decimal | -43.921.363 |
| geolocation_city | Ciudad | Texto | belo horizonte |
| geolocation_state | Estado | Texto | MG |
| | | | |
| Dataset | Marketing | | |
| Columna | Detalle | Tipo de dato | Ejemplo |
| mql_id | ID cliente potencial | Alfanumerico | 434c2eb8627ed4e1a0a4f0ee5d6022aa |
| first_contact_date | Fecha primer contacto | Fecha | 2018-03-22 |
| landing_page_id | informacion incompleta | Alfanumerico | 0d6bc3c00e4e64927cae2e8d9c6a0b9b |
| origin | Origen contacto | Texto | paid_search |
| | | | |
| Dataset | Order_items | | |
| Columna | Detalle | Tipo de dato | Ejemplo |
| order_id | ID orden de compra | Alfanumerico | ec1ecce6ed2f4a351a045a8de255e3af |
| order_item_id | ID identificacion articulo | Numerico entero | 1 |
| product_id | ID producto | Alfanumerico | 35afc973633aaeb6b877ff57b2793310 |
| seller_id | ID vendedor | Alfanumerico | 4a3ca9315b744ce9f8e9374361493884 |
| shipping_limit_date | Fecha limite de envio | Fecha y hora | 2017-09-21 10:10:20 |
| price | Precio | Numerico decimal | 89.90 |
| freight_value | Precio flete | Numerico decimal | 14.95 |
| | | | |

| Dataset | Order_payments | | |
|---|---|---|---|
| Columna | Detalle | Tipo de dato | Ejemplo |
| order_id | ID orden de compra | Alfanumerico | 647255bbedcdf748e7496180374b0dfe |
| payment_sequential | secuencia unificacion de medios de pago | Numerico entero | 1 |
| payment_type | Tipo de medio de pago | Texto | credit_card |
| payment_installments | numero de cuotas | Numerico entero | 1 |
| payment_value | valor del pago | Numerico decimal | 28.09 |
| | | | |
| Dataset | Order_reviews | | |
| Columna | Detalle | Tipo de dato | Ejemplo |
| review_id | ID reseña | Alfanumerico | a51b17cc0ae35deec0466cbf057b6700 |
| order_id | ID orden de compra | Alfanumerico | d3d9fd64df1cf1428ea9f88aaabb4713 |
| review_score | Calificacion reseña | Numerico entero | 3 |
| review_comment_title | Titulo reseña | Texto | Talvez recomendaria |
| review_comment_message | Mensaje reseña | Texto | A empresa deveria responder aos e-mail dos cli... |
| review_creation_date | Fecha envio reseña | Fecha y hora | 2018-05-17 0:00:00 |
| review_answer_timestamp | Fecha respuesta reseña | Fecha y hora | 2018-05-21 17:40:50 |
| | | | |

| Dataset | Orders | | |
|---|---|---|---|
| Columna | Detalle | Tipo de dato | Ejemplo |
| order_id | ID orden de compra | Alfanumerico | bb01789fc0271409b394c7b283e9dd29 |
| customer_id | ID cliente | Alfanumerico | ac3e01509ab5a9e7bb1bb344048ef81d |
| order_status | Estado pedido | Texto | delivered |
| order_purchase_timestamp | fecha compra | Fecha y hora | 2018-08-07 12:08:49 |
| order_approved_at | fecha aprobacion compra | Fecha y hora | 2018-08-07 12:24:47 |
| order_delivered_carrier_date | fecha entrega socio logistico | Fecha y hora | 2018-08-09 14:22:00 |
| order_delivered_customer_date | fecha entrega pedido | Fecha y hora | 2018-08-14 22:13:42 |
| order_estimated_delivery_date | fecha estimada de entrega informada | Fecha y hora | 2018-08-28 0:00:00 |
| | | | |
| Dataset | Products | | |
| Columna | Detalle | Tipo de dato | Ejemplo |
| product_id | ID producto | Alfanumerico | 093f7389fa2eccda5e86add8da4aa19e |
| product_category_name | nombre categoria producto | Texto | cama_mesa_banho |
| product_name_lenght | longitud del nombre del producto | Numerico decimal | 40.0 |
| product_description_lenght | longitud de la descripcion del producto | Numerico decimal | 718.0 |
| product_photos_qty | cantidad de fotos del producto | Numerico decimal | 1.0 |
| product_weight_g | peso del producto en gramos | Numerico decimal | 2000.0 |
| product_length_cm | longitud del producto en cm | Numerico decimal | 38.0 |
| product_height_cm | altura del producto en cm | Numerico decimal | 20.0 |
| product_width_cm | ancho del producto en cm | Numerico decimal | 25.0 |
| | | | |
| Dataset | Sellers | | |
| Columna | Detalle | Tipo de dato | Ejemplo |
| seller_id | ID vendedor | Alfanumerico | 282c7480173bb9c01dd41cc739fec010 |
| seller_zip_code_prefix | codigo postal vendedor | Numerico entero | 4795 |
| seller_city | ciudad del vendedor | Texto | sao paulo |
| seller_state | estado del vendedor | Texto | SP |
| | | | |
| Dataset | Product_category_name_translation | | |
| Columna | Detalle | Tipo de dato | Ejemplo |
| product_category_name | nombre categoria producto | Texto | moveis_quarto |
| product_category_name_english | nombre categoria producto en ingles | Texto | furniture_bedroom |
| | | | |
| Dataset | Valoracion | | |
| Columna | Detalle | Tipo de dato | Ejemplo |
| seller_id | ID vendedor | Numerico entero | 629 |
| seller_city | Ciudad vendedor | Texto | maua |
| seller_state | Estado vendedor | Texto | SP |
| distinct_prod | ID producto | Numerico entero | 58 |
| delivery_avg | promedio de envios | Numerico decimal | -6.981.012 |
| review_avg | promedio de reseñas | Numerico decimal | 4.431.034 |
| total_orders | ID orden | Numerico entero | 58 |
| total_income | precio total | Numerico decimal | 2548.80 |

## Informe preliminar de calidad del dato

Se importaron los datasets de Olist los cuales son de acceso público. Los datos constan de 11 archivos .csv de los cuales se utilizaron 10. Estos archivos se encuentra en la carpeta "Datasets/Datasets_original”

## Errores encontrados

A continuación se detallan los errores encontrados en el análisis exploratorio de los datos de cada dataset

• **olist_customers_dataset**

Esta tabla contiene 5 columnas, con 99441 valores en cada una. Faltan datos de customers_unique_id que no hay referencia de cómo fueron creados esos codigos. Esta tabla se puede utilizar para saber de qué ciudades provienen los consumos. No contiene datos nulos.

• **olist_geolocation_dataset**

Esta tabla cuenta con 5 columnas y 1000163 valores en cada una. Esta tabla puede servir para ubicar las ciudades correctamente en un mapa del Dashboard. No contiene datos nulos.

• **olist_order_items_dataset**

Esta tabla contiene 7 columnas con 112650 valores en cada una. Se puede utilizar para conocer cuál es el producto más vendido como así también cuánto facturó, cuál es el vendedor con más órdenes y por cuanto. No contiene valores nulos. Presenta outliers.

**• olist_order_payments_dataset**

Esta tabla cuenta con 5 columnas y 103886 valores en cada una. Se podría averiguar el comportamiento de los pagos. La utilidad dependerá de la propuesta del PO. No contiene valores nulos.

• **olist_order_reviews_dataset**

Esta tabla cuenta con 7 columnas y 99224 valores en cada una. La utilidad de esta tabla dependerá del objetivo propuesto. Contiene valores nulos en dos columnas.

• **olist_orders_dataset**

Esta tabla contiene 8 columnas y 99441 valores en cada una. La utilidad de esta tabla dependerá del objetivo. Contiene pocos valores nulos en 3 columnas.

• **olist_products_dataset**

Esta tabla contiene 9 columnas y 32951 valores en cada una. La utilidad de esta tabla dependerá del objetivo. Presenta muy pocos valores nulos en la mayoría de las columnas.

• **olist_sellers_dataset**

Esta tabla contiene 4 columnas con 3095 valores en cada una. No presenta datos nulos. Se podría conocer cuál es la ciudad con mayor cantidad de órdenes como así también el monto mayor en ventas.

• **product_category_name_translation**

Esta tabla contiene 2 columnas con 71 valores en cada una. No presenta errores.

• **olist_marketing_qualified_leads_dataset**

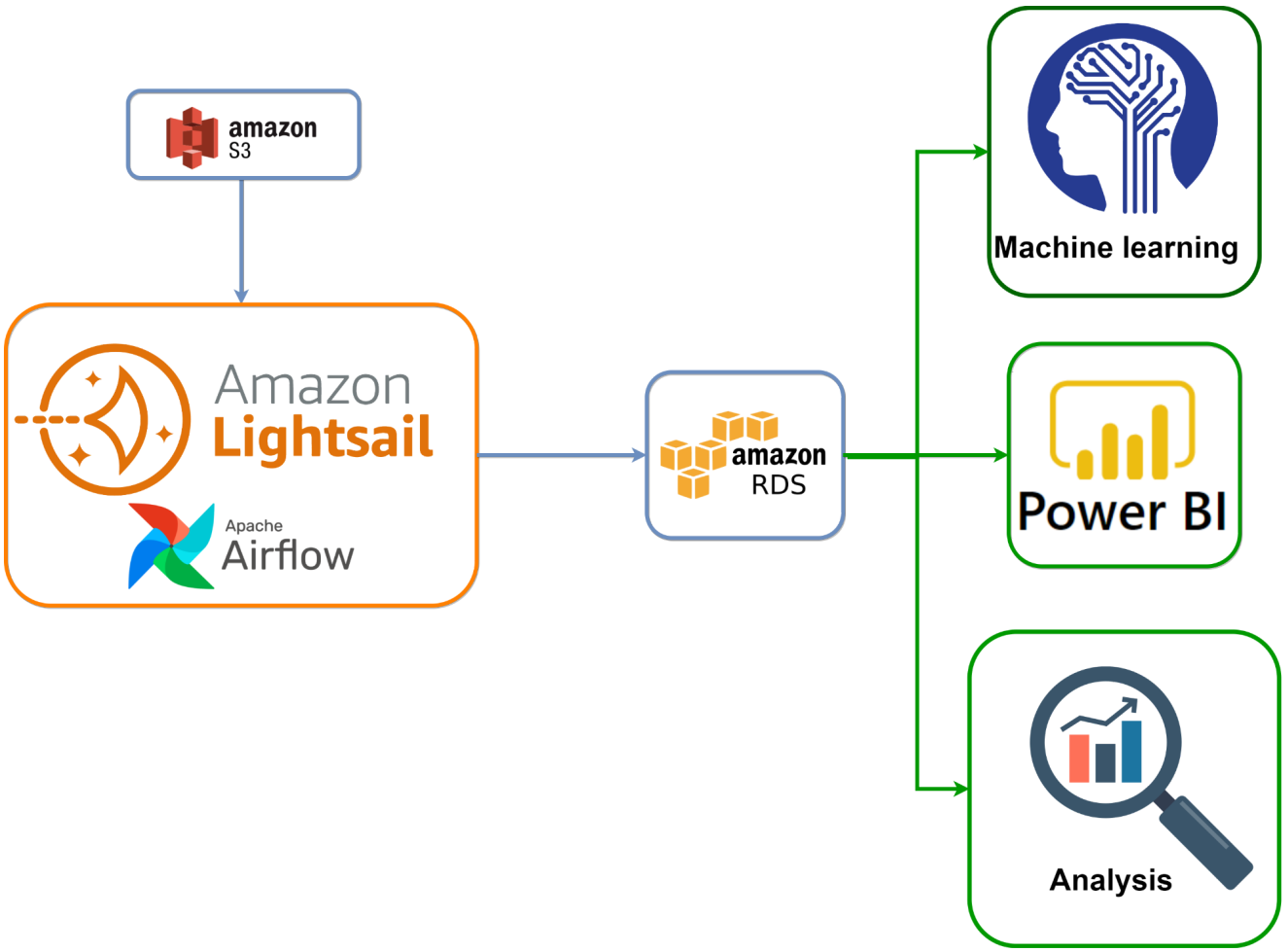
Esta tabla contiene 4 columnas, con un total de 8000 valores en cada una de ellas. Esta tabla no aporta datos significativos para el análisis que queremos realizar. Contiene pocos valores nulos en una columna.

• **olist_closed_deals_dataset**

Esta tabla consta de 14 columnas de datos, con un total de 842 valores en cada una de ellas. Las columnas has_company, has_gtin, average_stock y declared_product_catalog_size tienen más del 90 % de sus datos faltantes. Esta tabla no aporta datos muy relevantes. No se encontraron registros duplicados. Los nombres de los campos se encuentran en inglés.

# Extracción, transformación y carga de los datos

Una vez hecho el análisis preliminar de los datos, se procede a la transformación y carga de los mismos. Primero cargamos los datasets a trabajar en la nube usando el servicio de Amazon S3, luego diseñamos un script de python que contiene las funciones de carga y transformación y luego se utilizó airflow en una máquina virtual en la nube, utilizando el servicio Amazon Lightsail, para armar un pipeline él con este script de python. Elegimos configurar el pipeline para que se ejecute diariamente desde la nube, el cual va a tomar los datos del bucket de S3, los va a transformar y luego los va a cargar a la base de datos en amazon RDS



# Limpieza y normalización de los datos

A cada dataset se le eliminaron las filas duplicadas.

Al dataset “Closed_deals” se le cambió el tipo de datos en las columnas 'has_gtin' y 'has_company' al tipo de dato float64.

Al dataset Orders se le agregó un nueva columna llamada “tiempo_entrega” la cual se obtiene calculando la diferencia en días de las columnas “order_approved_at” y “order_delivered_customer_date”.

La transformación principal se llevó a cabo al combinar los datasets “orders”, ”order_reviews”, ”order_payments”, ”order_items”, “sellers” y “products” en una nueva tabla llamada datasets_combinados.

A esta tabla combinada se le creó una nueva columna llamada avg_income_month la cual contiene el cálculo del promedio de ingresos mensuales de cada vendedor.

Luego se crea una nueva tabla llamada Valoration y se toman las columnas “seller_id”, “seller_city”, “product_id”, “product_category_name”, “tiempo_entrega”, “review_score”, “order_id”, “payment_value” y “avg_income_month” del dataframe “datasets_combinados”.

Se realizan los respectivos cálculos y se cambian los siguiente nombres de columnas:
'product_id'→'distinct_prod'
'Tiempo_entrega'→'delivery_avg'
'product_category_name'→'distinct_categories'
'review_score'→'review_avg'

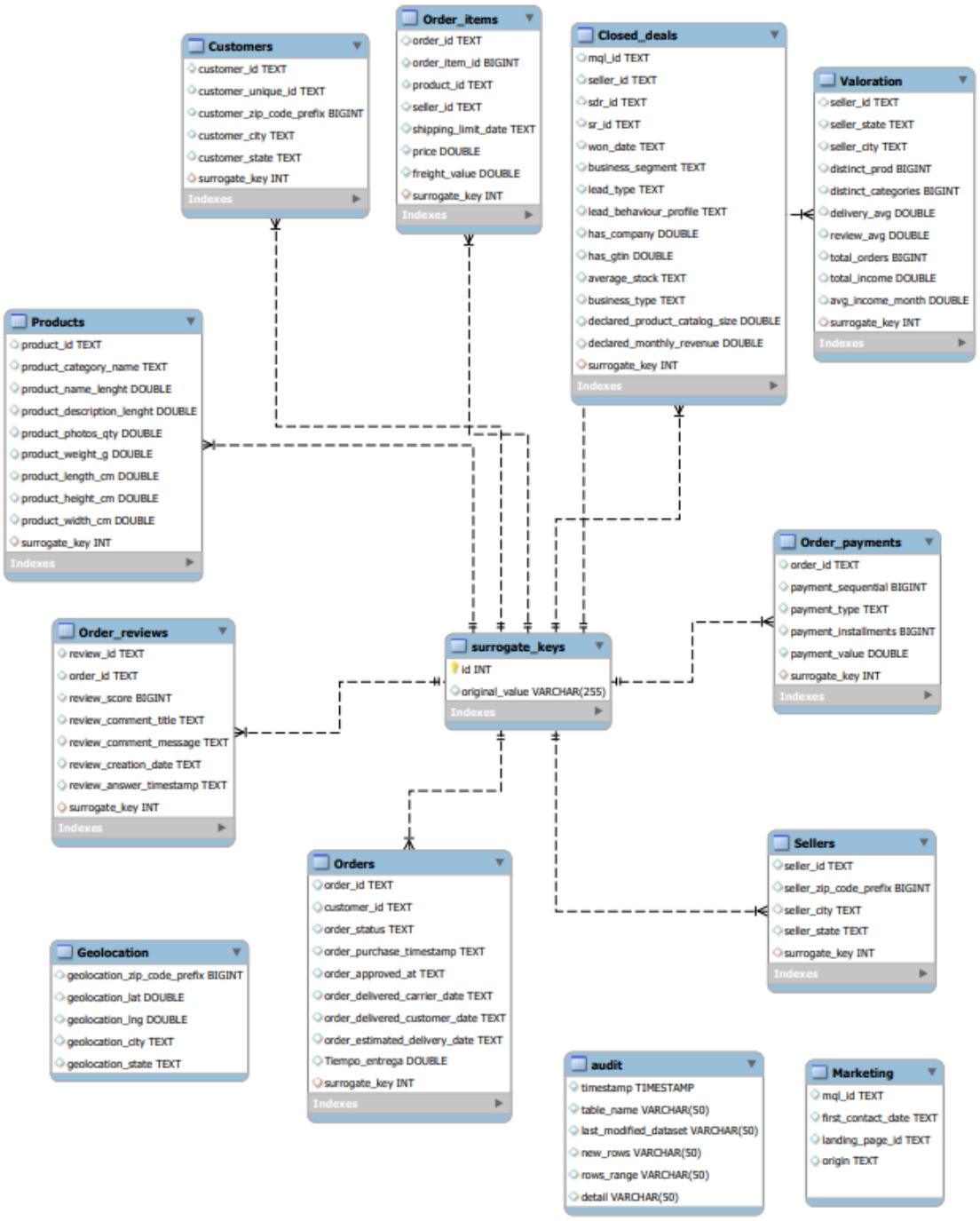
'order_id'→'total_orders'
'payment_value'→'total_income'

Las filas con valores de delivery_avg mayores a 50 días se eliminaron al ser pocas filas y considerarse outliers, Luego se eliminaron 125 filas que contengan valores nulos en alguna columna.

Por último se crearon funciones que dividen la tabla de valoración en distintas categorías y asignan distintos score a los vendedores en base a distintas métricas. Se agregaron estos valores en 5 nuevas columnas:

- **tier** = Esta columna asigna una categoría al vendedor
- **total_income_kpi**= Calculada en base a la columna “total_income” que contiene los ingresos totales del vendedor y segun la categoria del vendedor
- **delivery_avg_kpi**= Calculada según la columna delivery_avg y según la categoría asignada al vendedor.
- **review_avg_kpi**= Se calcula con la columna review_avg la cual contiene el promedio de review_score del vendedor y en base a la categoría asignada al vendedor.
- **performance_score**= Es un promedio de los tres kpi calculados anteriormente y se toma como el score final que se le asigna al vendedor para evaluar su rendimiento

# Modelo entidad-relación




Para el modelo de entidad relación se creó una tabla de claves subrogadas, la cual asigna a los ID de las distintas tablas, los cuales tienen valores alfanuméricos, un nuevo número entero como ID en una nueva columna para cada tabla llamada subrrogate_key.

Esto nos permite conectar todas las tablas y a su vez mantener los ID originales, los cuales están encriptados, de forma que el cliente puede volver a acceder a esos datos en caso de necesitarlo desencriptando los valores del ID.

A su vez diseñamos el sistema de carga automática para ser compatible con una carga incremental de datos. De esta forma, al actualizar los datasets del bucket de S3 el pipeline de airflow automáticamente cargará solo las nuevas filas a la base de datos, evitando así cargar filas repetidas. El pipeline a su vez comprobará la última fecha de modificación de los datasets cargados en la nube, de manera que solo va a operar con los que presenten modificaciones de la última fecha de modificación de la que se tenga registros, para cada dataset en específico.

Por último el pipeline de airflow va a cargar informacion en una tabla de auditoria, la cual contienen las siguientes columnas:

- **Timestamp=** fecha y hora en la que se realizó la operación en la base de datos.
- **table_name=** nombre de la tabla en la que se realiza la operación.
- **last_modified_dataset=** última fecha de modificación del archivo .csv que se encuentra en el bucket de S3, de la tabla con la que se está operando.
- **new_rows=** número de nuevas filas agregadas a la tabla
- **rows_range=** índice del rango de nuevas filas agregadas a la tabla.

# Conclusiones

Luego de analizar la información recabada, podemos inferir que:

✓ Los problemas en las entregas de los productos pueden ser causantes de valoraciones negativas por parte de los clientes a sus experiencias de compra.

✓ El mayor porcentaje de puntuaciones bajas procede de los clientes pertenecientes a las regiones Nordeste y Norte de Brasil.

✓ Las 5 ciudades con porcentajes más altos de bajas valoraciones son São Paulo; Rio de Janeiro; Belo Horizonte; Brasilia y Curitiba. ✓ Las categorías de productos peores puntuadas son “cama-mesa banho”; “beleza-saude”; “informática-accesorios”; “moveis decoracao” y “Esporte-lazer”.

✓ Se registran demoras en las entregas de los productos.

✓ Las regiones con mayor demora son la región Norte y la región Sur.

✓ En el caso de la región Norte, se observa una coincidencia entre la presencia de demoras en la entrega de productos y el elevado número de puntuaciones negativas registradas por los clientes.

✓ El costo promedio del flete es más elevado en la zona Norte y más bajo en la zona Sudeste.

✓ La ausencia de fotografías del producto a vender, incide en la menor cantidad de compras del mismo.

✓ La mayor cantidad de ventas se registran en el período que va desde el mes de marzo hasta el mes de agosto.

✓ Agosto es el mes de más ventas y septiembre su contraparte. ✓ La franja horaria más activa para las compras va desde las 10 AM a las 9 PM.

✓ Las 5 categorías de productos con mayor cantidad de ventas son cama_mesa_banho; beleza_saude; esporte_lazer; informatica_acessorios y moveis_decoracao y las 5 con menor cantidad son cds_dvds_musicais; la_cuisine;

pc_gamer;  fashion_roupa_infanto_juvenil y seguros_e_servicos.

✓ La región con mayor porcentaje de ventas es la región Sudeste y la  que menos ventas registra es la región Norte.

# Recomendaciones