

**Índice**

[**Introducción**](#_heading=h.gjdgxs) **2**

[**Infraestructura del proyecto - Workflow**](#_heading=h.30j0zll) **2**

[**Análisis exploratorio y calidad de datos**](#_heading=h.b9vajbunxxa2) **4**

[Exploratory Data Analysis](#_heading=h.n4qwtspo9ock) 4

[Calidad de los datos](#_heading=h.txn4i82l4leb) 4

[Decisión sobre transformaciones y acciones](#_heading=h.qdwkj6l9mq5c) 5

[**DataWarehouse y DER**](#_heading=h.ih8z3db64g3y) **6**

[**Pipeline ETL (Extract, Transform, Load)**](#_heading=h.1t3h5sf) **8**

[Extracción](#_heading=h.4d34og8) 8

[Transformación](#_heading=h.2s8eyo1) 8

[Carga](#_heading=h.fmcwridfyxm9) 8

[Carga inicial](#_heading=h.huxiwvbjmnom) 8

[Carga Incremental](#_heading=h.kuhtgonc6x18) 9

[**Conclusión**](#_heading=h.3rdcrjn) **9**

# Introducción

La presente documentación tiene como objetivo detallar las tareas y actividades realizadas durante la semana 2 del proyecto de análisis para el ecommerce Olist.

Esta semana el foco estuvo puesto en la ingeniería de datos, es decir, se realizó la planificación, diseño y construcción de la infraestructura del proyecto.

Las principales tareas se centraron en: análisis exploratorio de los datos, validación de la calidad de los mismos y por último la creación de pipelines para realizar el proceso de ETL con el objetivo de almacenar los datos en una estructura de tipo Data Warehouse.

Por último recordamos que los detalles de todo el análisis, documentación, datos y código se encuentran en el repositorio de Github: [Link](https://github.com/MelinaRG/Proyecto-Final-DATA)

# **Infraestructura del proyecto - Workflow**

Cuando hablamos de infraestructura, nos referimos al conjunto de elementos, herramientas (stack tecnológico) y procesos necesarios para almacenar, procesar y analizar los datos, con el fin de garantizar que el proyecto se ejecute de manera eficiente y segura.

En relación a las **herramientas** y **sistemas** elegimos:

* Fuente de datos: Kaggle
* Sistema de almacenamiento de datos (Datawarehouse), utilizando PostgreSQL
* Herramientas de procesamiento de datos: Python (Pandas, Numpy, Matplotlib, Seaborn)
* Herramientas de análisis de datos: Sklearn para el desarrollo de modelos de ML.
* Herramientas de visualización de datos: Power Bi para la elaboración de un dashboard que cuente con métricas, gráficos y KPIs. Streamlit para la visualización del modelo de ML
* Sistemas de almacenamiento y procesamiento de datos (escalabilidad): Azure
* Sistemas de Automatización y monitoreo de datos: Airflow

En relación a los **procesos** que se llevarán encontramos:

* Recolección de datos: Esta etapa se encarga de obtener los datos necesarios para el proyecto
* Limpieza de datos: Esta se encarga de identificar y corregir problemas en los datos, como valores faltantes, duplicados, outliers o valores atípicos, para garantizar que los datos sean precisos y confiables.
* Integración de datos: Esta etapa se encarga de combinar los datos de diferentes fuentes en una sola estructura de datos, utilizando técnicas la normalización.
* Transformación de datos: Esta etapa se encarga de modificar los datos para que cumplan con las necesidades específicas del proyecto.
* Análisis de datos: utilizando técnicas estadísticas, de aprendizaje automático o de visualización de datos, para extraer insights y patrones interesantes.
* Presentación de resultados: Esta etapa se encarga de presentar los resultados del análisis de manera clara, concisa y accesible, utilizando herramientas de visualización de datos o informes.

# 

# **Análisis exploratorio y calidad de datos**

El análisis de calidad de los datos ayuda a identificar y corregir problemas, como valores faltantes, duplicados, outliers o valores atípicos, que pueden afectar negativamente la precisión y validez de los resultados del análisis.

Una vez que los datos están limpios y validados, se pueden realizar análisis más precisos y confiables, lo cual es esencial para la toma de decisiones en una organización.

## Exploratory Data Analysis

El objetivo principal de este tipo de análisis es obtener una comprensión general de los datos y sus características, antes de aplicar técnicas más avanzadas de análisis y modelado.

Se divide en varios pasos:

* Carga y limpieza de los datos
* Análisis estadístico descriptivo: en algunas tablas se realizaron estadísticas básicas como medias, mediana y distribuciones para tener una idea de cómo se distribuyen los datos.
* Análisis gráfico: se utilizaron gráficos para visualizar los datos y detectar patrones y tendencias.
* Análisis de la relación entre variables: se buscó analizar las relaciones mediante técnicas como la correlación o el análisis de componentes principales.
* Identificación de problemas: se buscó detectar problemas en los datos, como outliers o valores atípicos.

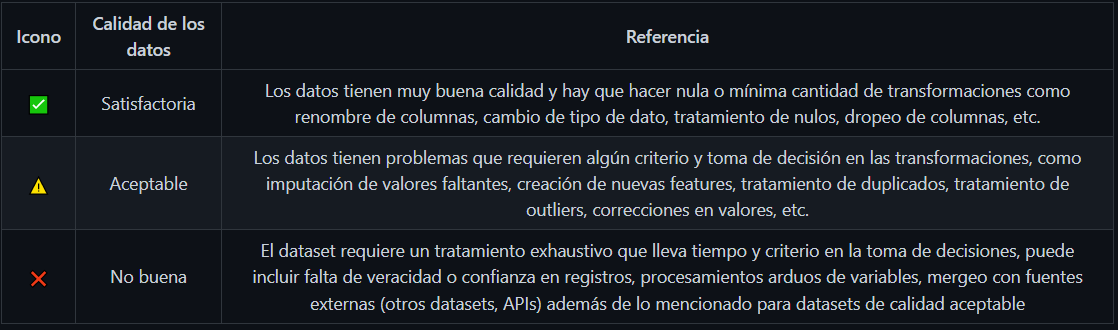
En este [Link](https://github.com/MelinaRG/Proyecto-Final-DATA/blob/EDA/Informe_EDA.md) puede acceder al Análisis exploratorio completo.

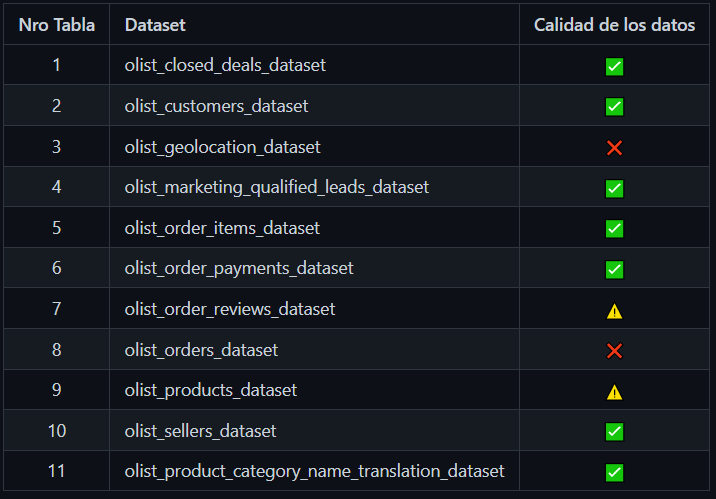
También se elaboró una tabla resumen de la calidad de cada dataset, se puede acceder a través del siguiente link

[Tabla Resumen Calidad de datos OLIST](https://docs.google.com/spreadsheets/d/1y7QMFwnjEGlktEs0kt9VLtGqG2xxH6sHzU4dBOnioVA/edit?usp=sharing)

## Calidad de los datos

Se decide un criterio de la calidad de datos del dataset analizado en forma particular luego de los análisis exploratorios correspondientes. El criterio es el siguiente:





Los datasets más comprometidos son los de órdenes y geolocalización, sobre los que se destina más tiempo y esfuerzo para lograr las transformaciones correspondientes.

## Decisión sobre transformaciones y acciones

De acuerdo a lo evaluado anteriormente se establecen puntos de transformación que se deben asentar en scripts para que luego sean fácilmente ejecutados en un pipeline las operaciones de extracción, transformación y carga. Las decisiones son:



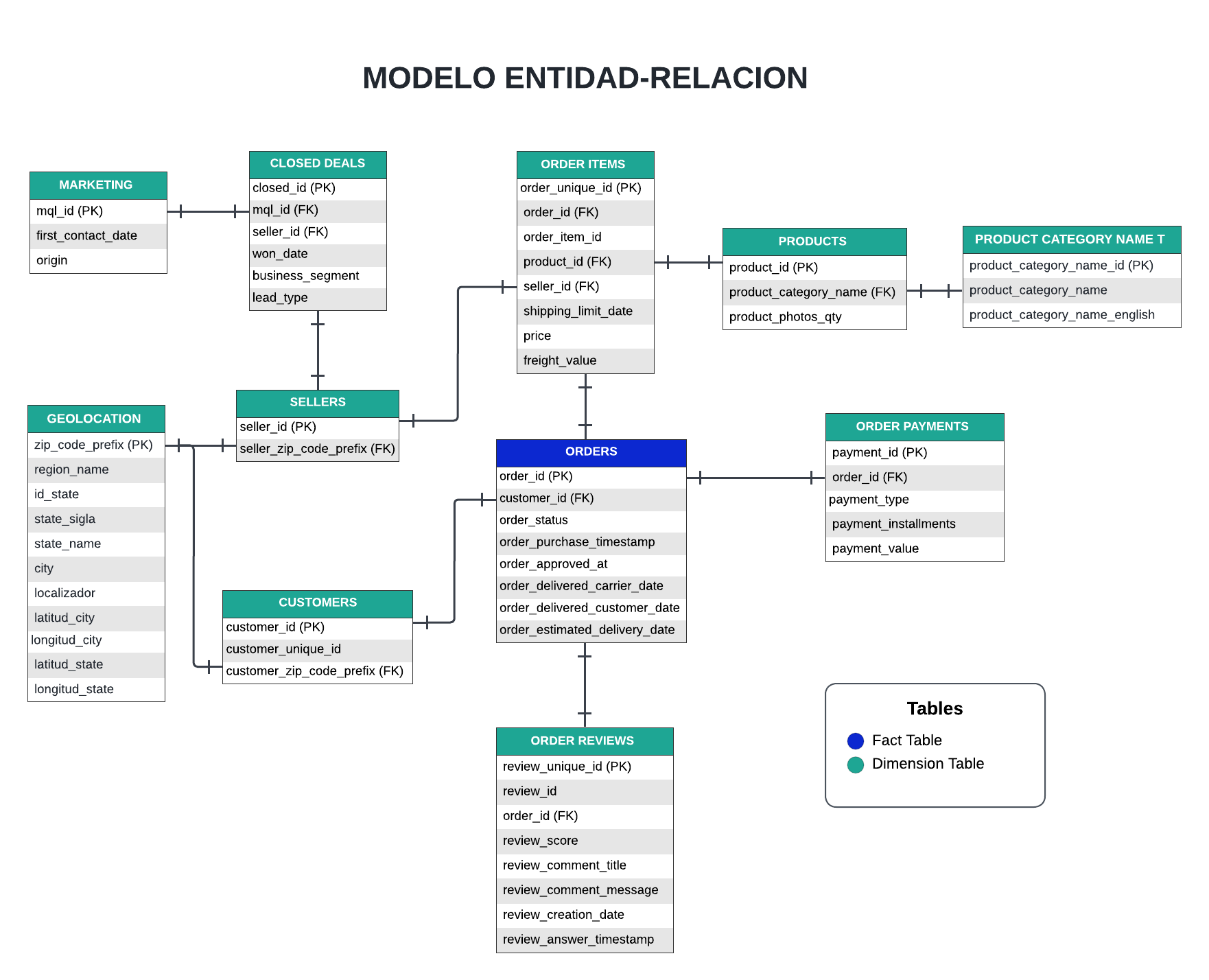
# **DataWarehouse y DER**

Para el data warehouse utilizamos Postgresql ya que es un sistema de gestión de bases de datos relacional con gran capacidad de escalabilidad, integridad referencial con la imposición de restricciones, buenas propiedades de seguridad y amplia compatibilidad con herramientas de análisis.

Empezamos definiendo el modelo entidad-relación (DER), que es una forma de modelar el mundo real en una base de datos de manera que sea fácil de comprender y utilizar para los usuarios.

Estructurando el modelo de Olist, encontramos que el mismo cuenta con:

* Tabla de hechos: 1
* Tabla de dimensiones: 10



Las tablas de **hechos** y las de **dimensiones** se relacionan mediante un proceso de diseño de base de datos conocido como modelado de estrella. Este modelo consta de una tabla central de hechos que contiene los datos medibles, en nuestro caso orders , y varias tablas de dimensiones que proporcionan contexto a esos datos.

Las tablas de dimensiones contienen información que describe un aspecto particular de los datos en la tabla de hechos, como el lugar, el producto o el vendedor, etc.

A su vez, estas tablas se relacionan entre sí por medio de las llamadas **claves primarias** y **claves foráneas**. Una clave primaria se utiliza para identificar de manera única una fila dentro de una tabla, generalmente son aquellas columnas que contienen id (ej. seller\_id, customer\_id, etc) en su nombre. Mientras que, una clave foránea se relaciona con una clave primaria de otra tabla, lo que garantiza que los datos de una tabla se refieren a registros existentes en la otra tabla.

# Pipeline **ETL (Extract, Transform, Load)**

Es un proceso automatizado que se utiliza para integrar, limpiar y transformar datos de diferentes fuentes y almacenarlos en una estructura de almacenamiento, como un Data Warehouse. Generalmente se divide en tres etapas:

## Extracción

En esta etapa se recolectaron los datos de diferentes fuentes, como archivos planos (.csv), APIs (geolacation), etc.

## Transformación

En esta etapa se limpian y transforman los datos para hacerlos coherentes y consistentes, y se preparan para su carga en el sistema de almacenamiento centralizado.

[Link](https://github.com/MelinaRG/Proyecto-Final-DATA/blob/EDA/Informe_EDA.md) al informe con las transformaciones detalladas tabla por tabla,

## Carga

### Carga inicial

Esta etapa se encarga de cargar los datos limpios y transformados en la estructura de almacenamiento final, en este caso el Datawarehouse.

Esta semana decidimos hacer el deploy de la base de datos en Render, ya que nos permitía a todos los integrantes ingresar a la misma, a través de las credenciales generadas.

Luego se desarrolló un script de python utilizando la librería *psycopg2* para la creación de las distintas tablas incluyendo cada columna con su tipo de dato como así también, incorporando en el mismo la clave primaria (PK) y las claves foráneas (FK).

Posteriormente, utilizamos la librería *sqlalchemy* para realizar el segundo script para la carga inicial de datos. Este Pipeline consta de varias funciones, una por cada tabla, dentro de las cuales se realiza el proceso de ETL: se extrae el dataset del archivo csv correspondiente, se realizan las transformaciones correspondientes y finalmente se realiza la carga de los datos procesados en el Datawarehouse.

### Escalabilidad : Carga Incremental

La idea para la próxima semana es implementar una carga incremental de los datos, con el objetivo de mejorar la escalabilidad del proyecto.

Para ello, vamos a utilizar un servidor cloud Azure, conectando con una base de datos Postgresql también alojada en la nube de Azure. Utilizaremos esta configuración para realizar la carga incremental y llevar a cabo el proceso ETL mediante un pipeline automatizado con Airflow.

# **Conclusión**

Con esta semana de ingeniería de datos, hemos concluido un importante paso en nuestro proyecto de análisis para el ecommerce Olist.

Hemos podido apreciar la importancia de este proceso en el proyecto, ya que permite evaluar la calidad de los datos y determinar si son confiables para llevar a cabo el análisis, además de evitar malgastar recursos en datos irrelevantes o outliers que podrían sesgar los resultados de nuestro análisis.

El próximo paso será analizar los datos en detalle en busca de patrones y tendencias, desarrollar modelos de predicción y generar un dashboard con los resultados y los indicadores clave de rendimiento (KPIs).