PROYECTO GRUPAL - HENRY

GRUPO 6 - SEMANA 2

E-COMMERCE - DATASET BY OLIST

JULIO 2022

ÍNDICE

EQUIPO DE TRABAJO

LEONARDO ARGAÑARAS

Periodista, estudiante de Data Science
Product Manager | Data Analist

JOAN DEMIAN GODOY

Licenciado y estudiante de Data Science.
Ingeniero de datos del proyecto.

EDGAR MORALES

Físico y Estudiante de Data Science en Henry.
Científico de Datos en el Proyecto.

KAREN BANEGAS

Ingeniera en Sistemas. Estudiante de Data Science.
Ingeniera de datos en el proyecto.

FACUNDO OPPIDO

Estudiante de Ing. en Sistemas, alumno de SoyHenry en la carrera de Data Science.
Ingeniero de datos del proyecto.

INTRODUCCIÓN

A continuación, se presentará el trabajo realizado en la semana 2 del proyecto Olist. Donde como equipo de trabajo, nos organizamos para poder llevar a cabo la implementación de un flujo de trabajo que sea capaz de cubrir las necesidades del cliente.

Este flujo de trabajo comprende desde la normalización de las tablas propuestas, hasta la creación de un modelo relacional (Data Warehouse) el cual ya incluye la ingesta de datos anteriormente ya pre-procesados.

OBJETIVO

El objetivo del trabajo de esta semana se basa en poder ya tener los datos procesados y almacenados en un Data Warehouse con un estructura adecuada para su manipulación en un servicio de análisis como lo es PowerBI.

En pocas y otras palabras el objetivo fue lograr la limpieza de los datos proporcionados para llevarlos a una estructura formal, la cual nos permitirá en un futuro usar una herramienta de visualización, para así poder hacer los reportes necesarios.

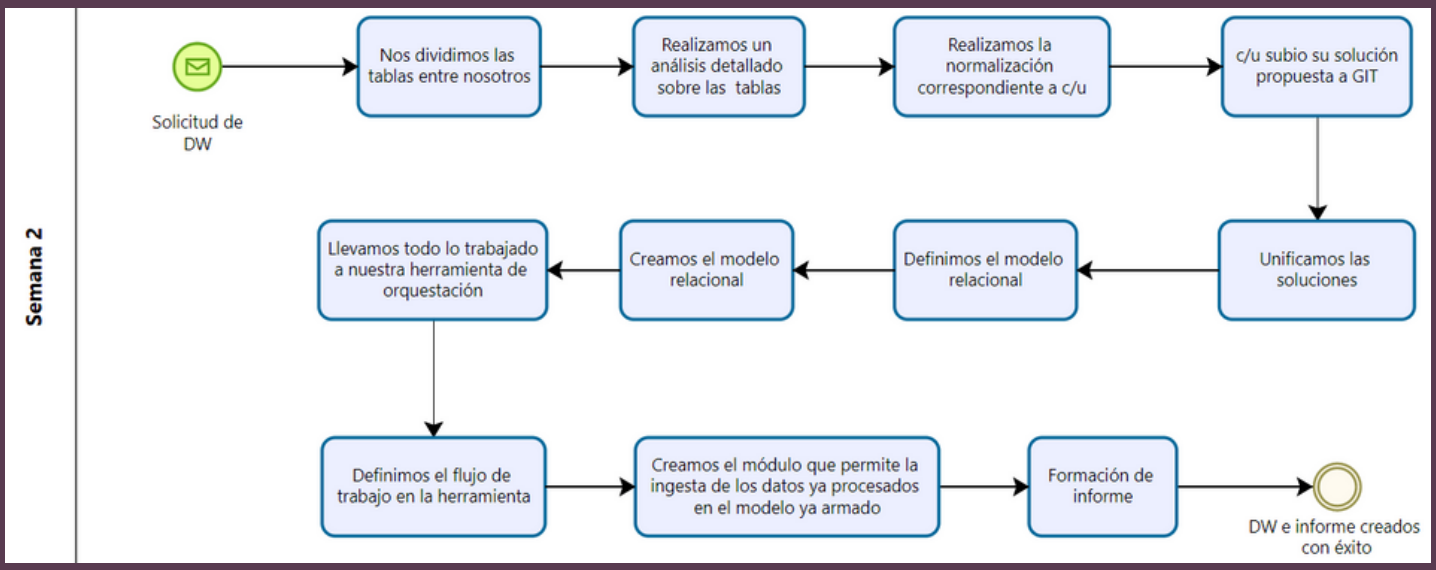
PLAN DE ACCIÓN

A continuación, se puede observar el conjunto de tareas que llevamos a cabo durante esta semana para poder lograr el objetivo de la misma.

Como evento disparador tenemos la solicitud de la implementación del Data Warehouse y como evento de fin tenemos el mismo ya generado con su informe respectivo.

El detalle de cada tarea es explicado por los integrantes del grupo de serlo necesario.

SECUENCIA DE TAREAS LLEVADAS A CABO



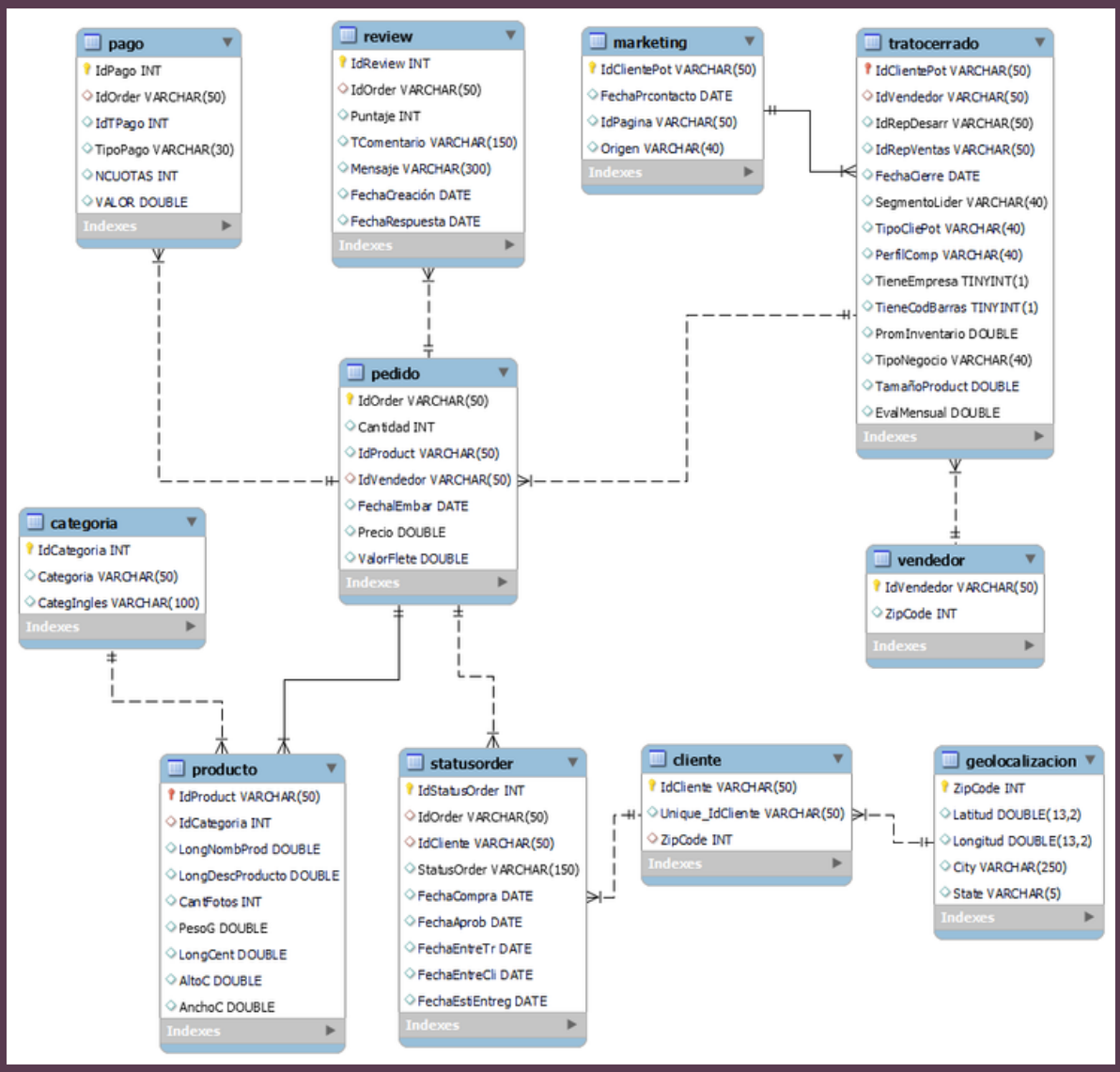
EJECUCIÓN DEL PLAN

A continuación, detallamos algunos puntos importantes sobre la ejecución del plan planteado teniendo en cuenta algunos detalles más técnicos.

NORMALIZACIÓN

- Respecto a los datos nulos encontrados en las tablas, se decidió reemplazarlos por la cadena de texto 'Sin Dato' en el caso de ser una columna de tipo String, y en el caso de columnas numéricas los valores como 'unknown' fueron reemplazados por NaN. Esto con el objetivo de poder manipular los campos numéricos sin tener problemas en los tipos de datos.
- Respecto a los valor atípicos identificados en las tablas solo se decidieron extraer aquellos que representaban un gran porcentaje, como por ejemplo en el caso de la columna 'declared_monthly_revenue' de la tablas *closed_deals.*
- Se identificaron registros con una localidad fuera del país de Brasil, a estos registros se decidió extraerlos por ser valor atípicos.
- Se cambiaron los nombres tanto de las columnas como el de las tablas para poder entender mejor el dominio. Básicamente se puso el mismo nombre pero en español.
- Todas las columnas que incluían un campo de fecha, venían como tipo de dato Object, las mismas fueron reemplazadas por tipo de dato DateTime, para su futura manipulación de serlo necesario.

CREACION DEL MODELO RELACIONAL



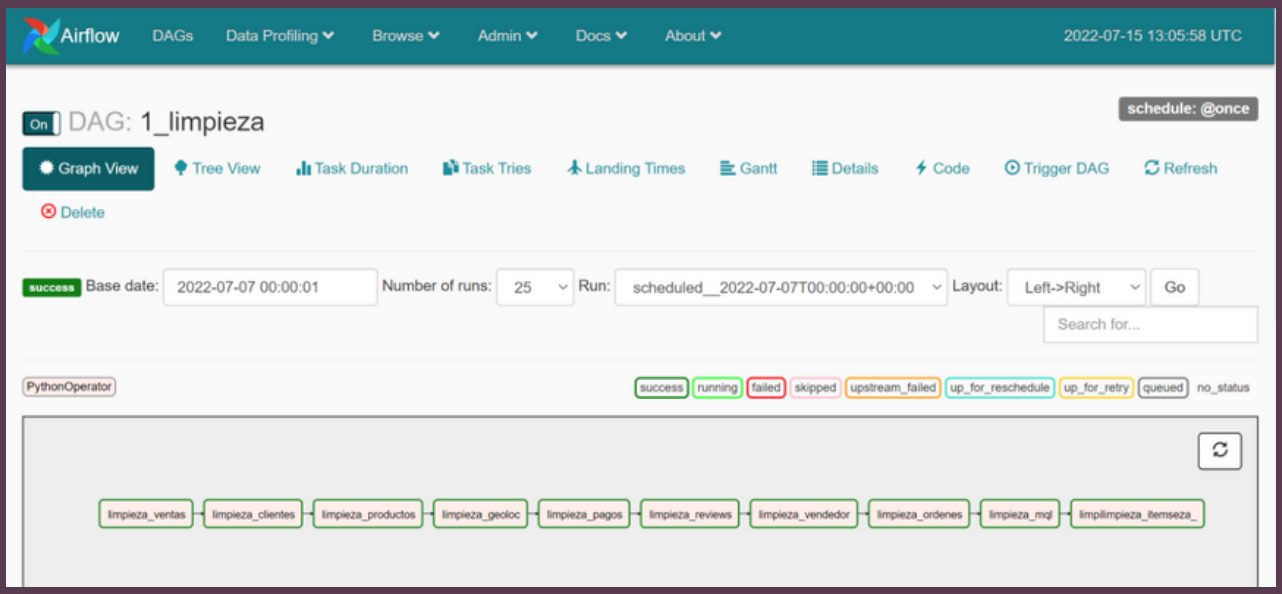
En cuanto a las relaciones que se llevaron a cabo para poder confeccionar el modelo relacional propuesto, se encuentran detalladas en el diccionario ubicado en el repositorio de GitHub.

Las mismas fueron elegidas gracias al pre-análisis confeccionado en la primer semana y con un poco más de detalle en el que se realizó esta misma semana.

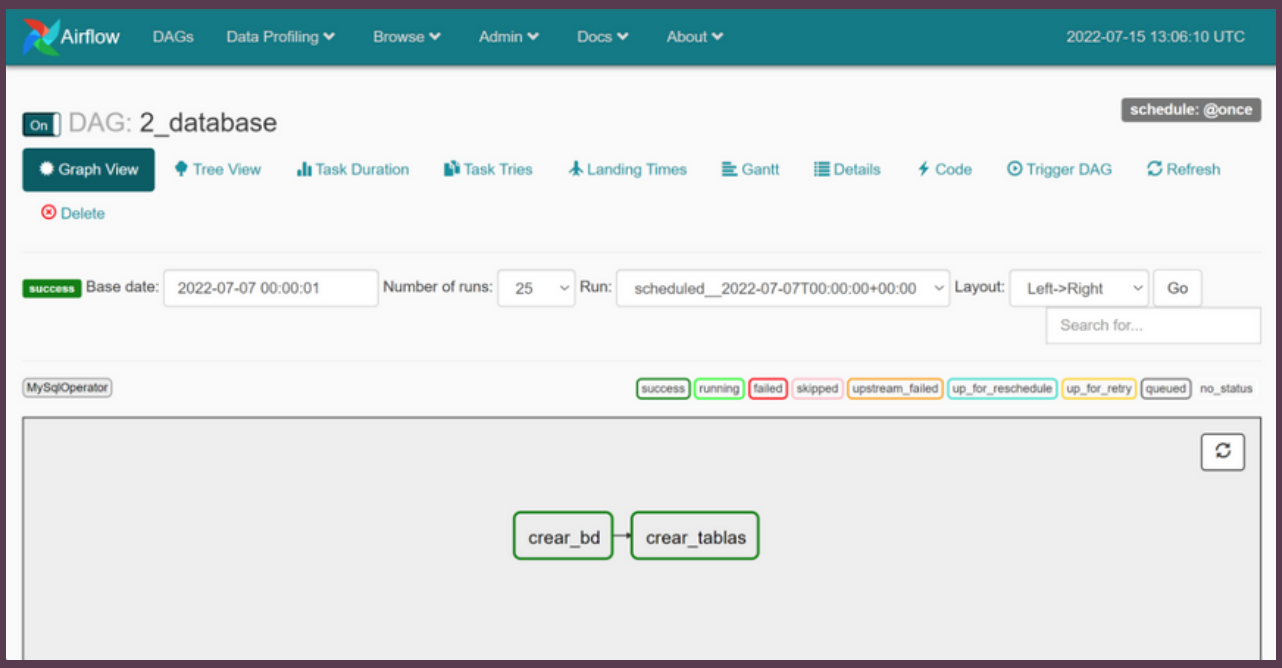
IMPLEMENTACIÓN DEL FLUJO DE TRABAJO EN AIRFLOW

En las siguientes imágenes se pueden observar las distintas secuencias de tareas, también conocido como flujo de trabajo, el cual fue organizado gracias a la herramienta de orquestación Airflow. Esta herramienta nos dió la posibilidad automatizar todo nuestro proceso, donde cada tarea o también denominado DAG que se logra ver, representa un conjunto de Scripts de código Python o SQL, dependiendo la tarea que se esté viendo.

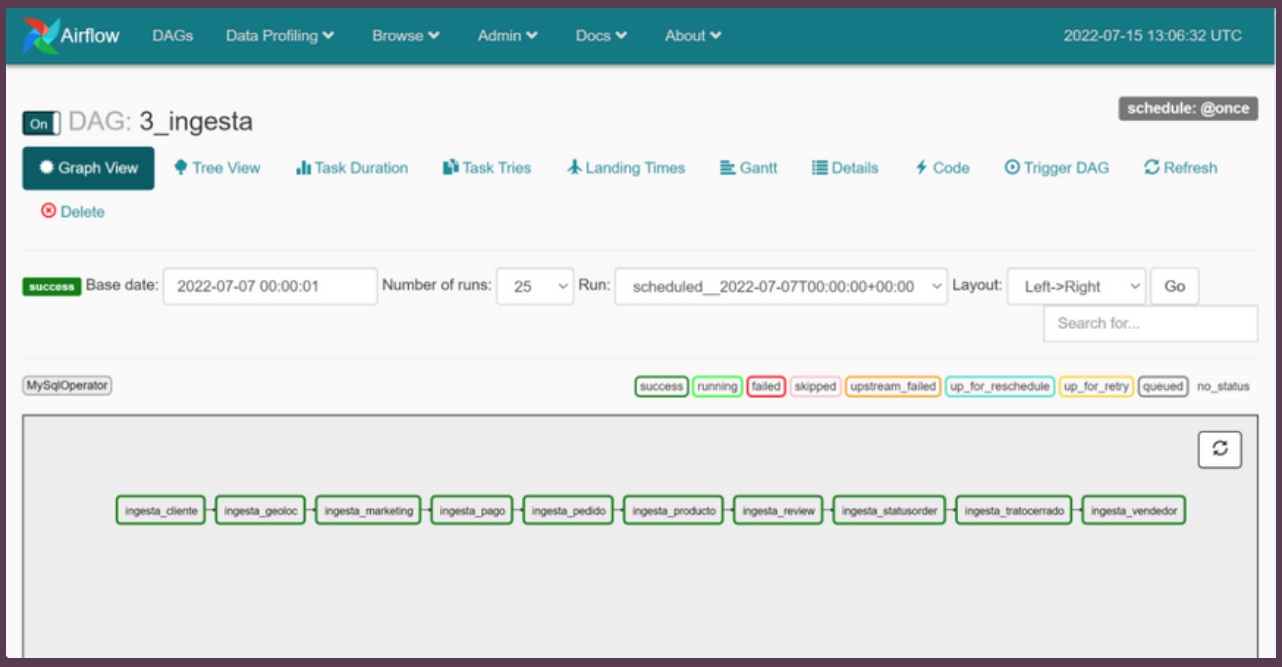
DAG DE NORMALIZACIÓN DE LAS TABLAS



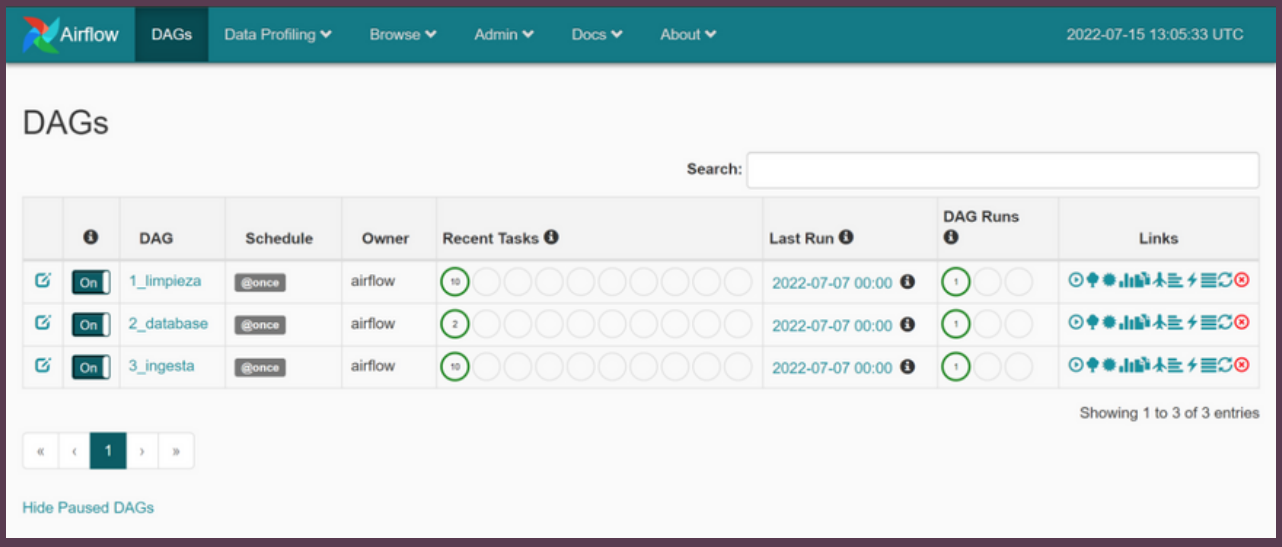
DAG DE CREACIÓN DE LA BASE DE DATOS Y DE LAS TABLAS



DAG DE INGESTA DE LOS DATOS EN CADA TABLA



CONJUNTO DE DAGS TRABAJADOS



REGLAS DE NEGOCIO

ID	Nombre	Descripción
1	Datos de tabla	Los datos de las tablas a trabajar no deben presentar problemas diferentes a los de las tablas presentadas desde un principio.
2	Formato del archivo	El sistema fue creado para trabajar con archivos que estén en formato (CSV).
3	Nombre del archivo	El nombre de los archivos a trabajar debe cumplir con el estándar marcado en la sección 1.0.
4	Ubicación de archivo	El sistema trabaja con rutas ya especificadas, por eso es necesario conocer donde se deben ubicar los datos crudos y donde se obtendrán los datos procesados, ver sección 1.1.
5	Totalidad de archivos	Deben ubicarse TODOS los archivos en la carpeta mencionada anteriormente.
6	Flujo de trabajo	El sistema funciona bajo un flujo de trabajo ya predefinido.

DETALLE DE REGLAS

SECCIÓN 1.0 - NOMBRE DEL ARCHIVO

Para que el sistema cumpla su objetivo con éxito es de suma importancia respetar el estándar del nombre del archivo, de lo contario el sistema no logrará reconocer que tipo de acción tomar sobre el archivo ingresado. A continuación se detallan los estándares a cumplir:

Nombre del archivo propuesto	Estándar a cumplir
olist_customers_dataset	olist_customers_dataset
olist_geolocation_dataset	olist_geolocation_dataset
olist_order_items_dataset	olist_order_items_dataset

Nombre del archivo propuesto	Estándar a cumplir
olist_order_payments_dataset	olist_order_payments_dataset
olist_order_reviews_dataset	olist_order_reviews_dataset
olist_orders_dataset	olist_orders_dataset
olist_products_dataset	olist_products_dataset
olist_sellers_dataset	olist_sellers_dataset
product_category_name_translation	product_category_name_translation
olist_closed_deals_dataset	olist_closed_deals_dataset
olist_marketing_qualified_leads_dataset	olist_marketing_qualified_leads_dataset

Es importante tener en cuenta que el estándar a cumplir debe incluirse solo y exclusivamente en el archivo indicado y no en otro, ya que sino, el sistema puede llegar a reconocer un archivo como otro que no debería ser.

SECCIÓN 1.1 - UBICACIÓN DEL ARCHIVO

Es muy importante verificar que antes de correr el flujo de trabajo, los archivos se encuentren en la ubicación correcta, a continuación se especifica la misma:
Los archivos CSV a ser procesados deben ubicarse en la carpeta 'import' con los nombres indicados con anterioridad. Los cambios reflejados sobre los mismos se verán en la carpeta 'limpia'.

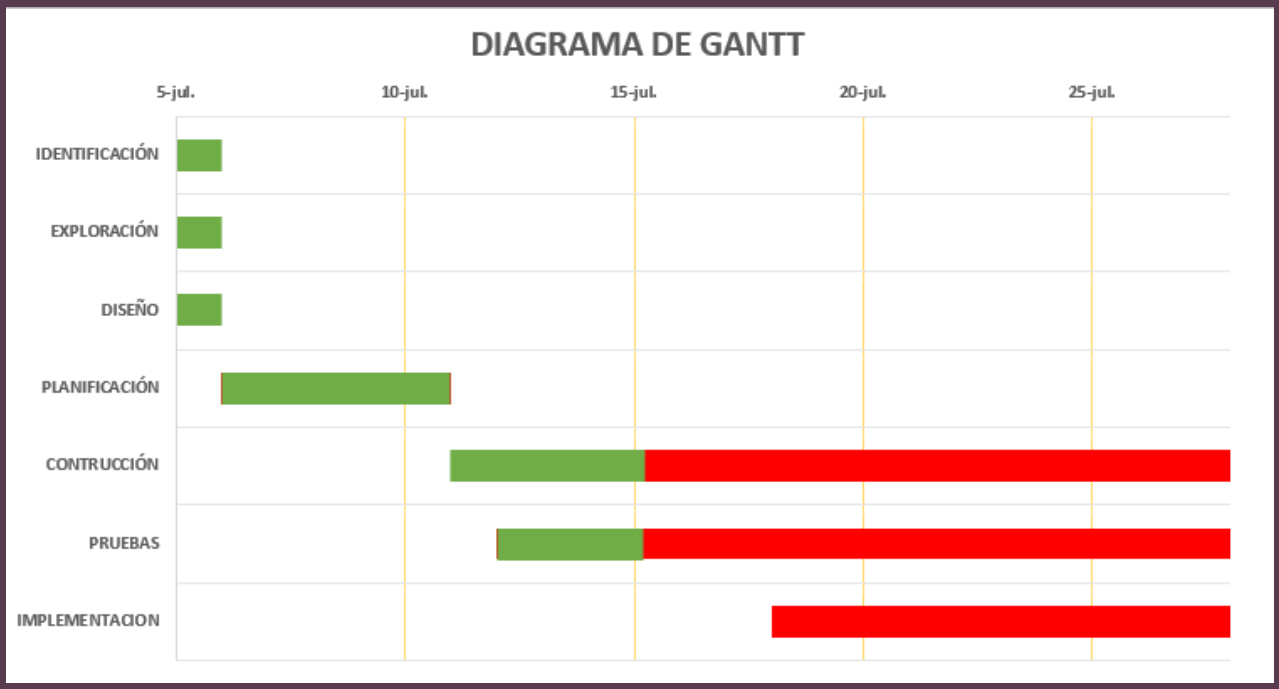
ORGANIZACIÓN

TRELLO

A continuación, se puede observar una captura de pantalla realizada en la página de Trello, donde estamos llevando a cabo nuestra organización de la segunda semana del proyecto. En ella se puede observar las diferentes tareas realizadas y quienes fueron los que la llevaron a cabo.

DIAGRAMA DE GANTT

ACTIVIDADES	FECHA INICIO	DURACION DIAS	FECHA FIN	% COMPLETADO	DIAS COMPLETADOS
IDENTIFICACIÓN	5-jul	1	6-jul	100%	1,00
EXPLORACIÓN	5-jul	1	6-jul	100%	1,00
DISEÑO	5-jul	1	6-jul	100%	1,00
PLANIFICACIÓN	6-jul	5	11-jul	100%	5,00
CONTRUCCIÓN	11-jul	17	28-jul	40%	6,80
PRUEBAS	12-jul	16	28-jul	30%	4,80
IMPLEMENTACION	18-jul	10	28-jul	0%	0,00



Utilizando el Excel creado en la primera semana podemos ir viendo el avance del progreso de nuestro proyecto. De esta forma, se puede medir la completitud de los procesos asociados.

STACK UTILIZADO

LINKS

Trello: https://trello.com/b/E6y2JA3A/grupo-6
GitHub: https://github.com/EdgarMH/DS_Grupo-6_Olist/

DEMO PRACTICA

Para finalizar se lleva a cabo una presentación de la demo sobre el trabajo realizado.