

# INFORME SPRINT Nro 1

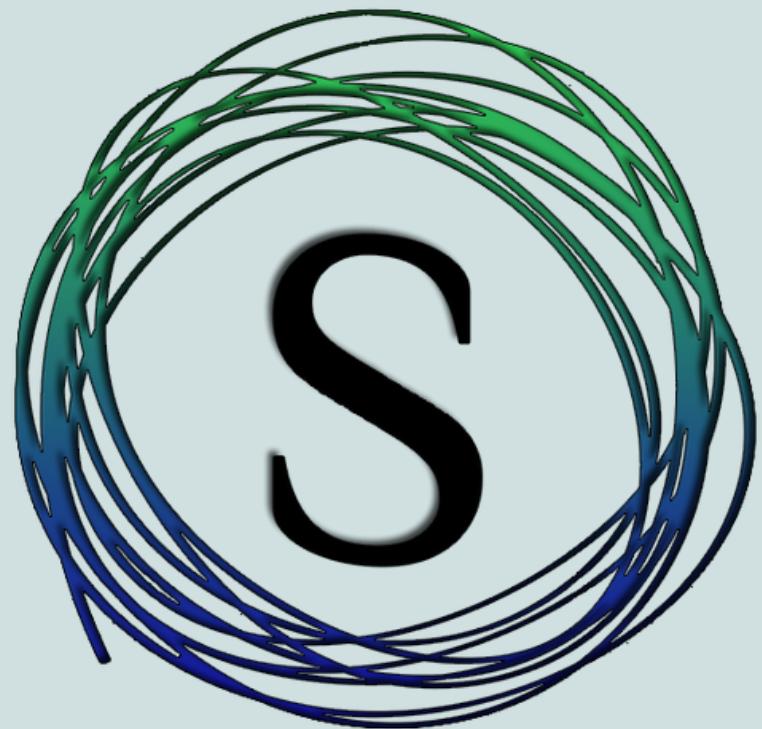
GRUPO 8



# ÍNDICE:

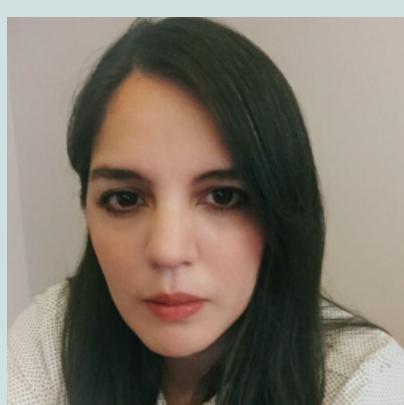
- 3- ¿Quiénes somos?
- 4- ¿Quién nos contrata?
- 5- Equipo de trabajo, roles y responsabilidades
- 6- Repositorio de GitHub
- 7- Cronograma general Gantt
- 8- Metodología de trabajo
- 9- EDA de los datos
- 10- Alcances y objetivos del proyecto:
- 11- KPI's
- 12- Stack tecnológico
- 13- Arquitectura

# ¿Quiénes somos?

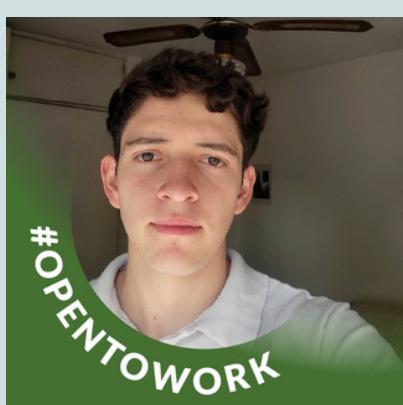


**Sinergia global**, una consultora especializada en el análisis de datos y estrategias empresariales.

Nuestra misión es ayudar a nuestros clientes para que extraigan el valor máximo de sus datos y posean las herramientas necesarias para tomar decisiones empresariales informadas y efectivas.



Edith Cuellar



Leandro Ibarra



Matias Baez



Nicolas Ibarra



Tinmar Andrade

# ¿Quién nos contrata?



**Culinary Cross Roads** es una empresa que gestiona varios establecimientos gastronómicos en el estado de Indiana. Tiene como objetivo expandirse a otros estados, y nos han confiado el análisis del mercado estadounidense actual.

Nuestra tarea principal consiste en proporcionar información valiosa basada en un análisis profundo de datos para respaldar decisiones estratégicas que minimicen los riesgos y maximicen las oportunidades de crecimiento para la empresa.

# Equipo de trabajo

## Roles y responsabilidades

<b>Data Science</b>	Edith Cuellar
<b>Data Analyst</b>	Leandro Ibarra
<b>Data Engineer</b>	Tinmar Andrade
<b>Machine Learning Engineer</b>	Matias Baez
<b>Analista y Coordinador</b>	Nicolas Ibarra

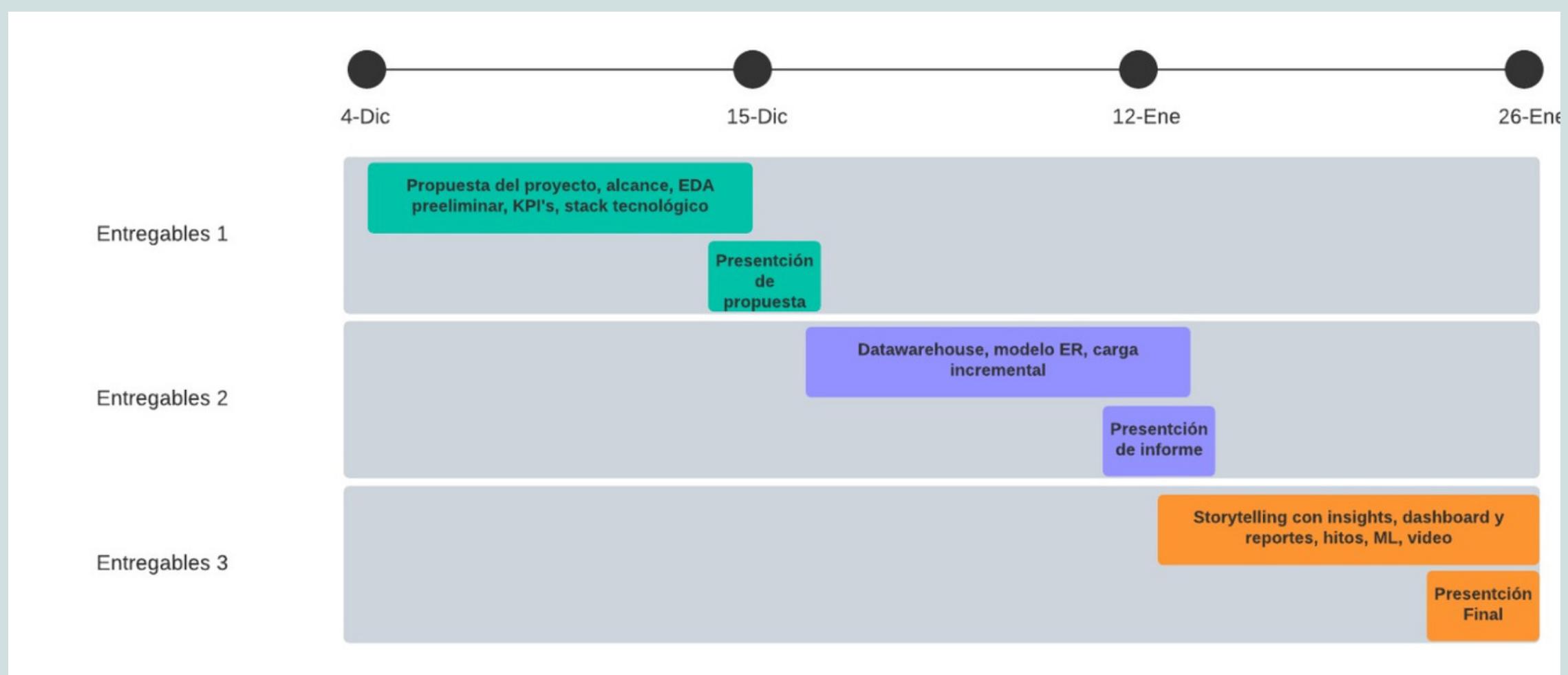
# Repositorio de GitHub

Para gestionar y compartir los avances del proyecto, hemos creado un repositorio en GitHub. Este espacio albergará todas las actualizaciones, cambios y el progreso general del desarrollo. También será el lugar donde se publicarán los resultados finales. Puedes acceder al repositorio haciendo clic en el siguiente enlace

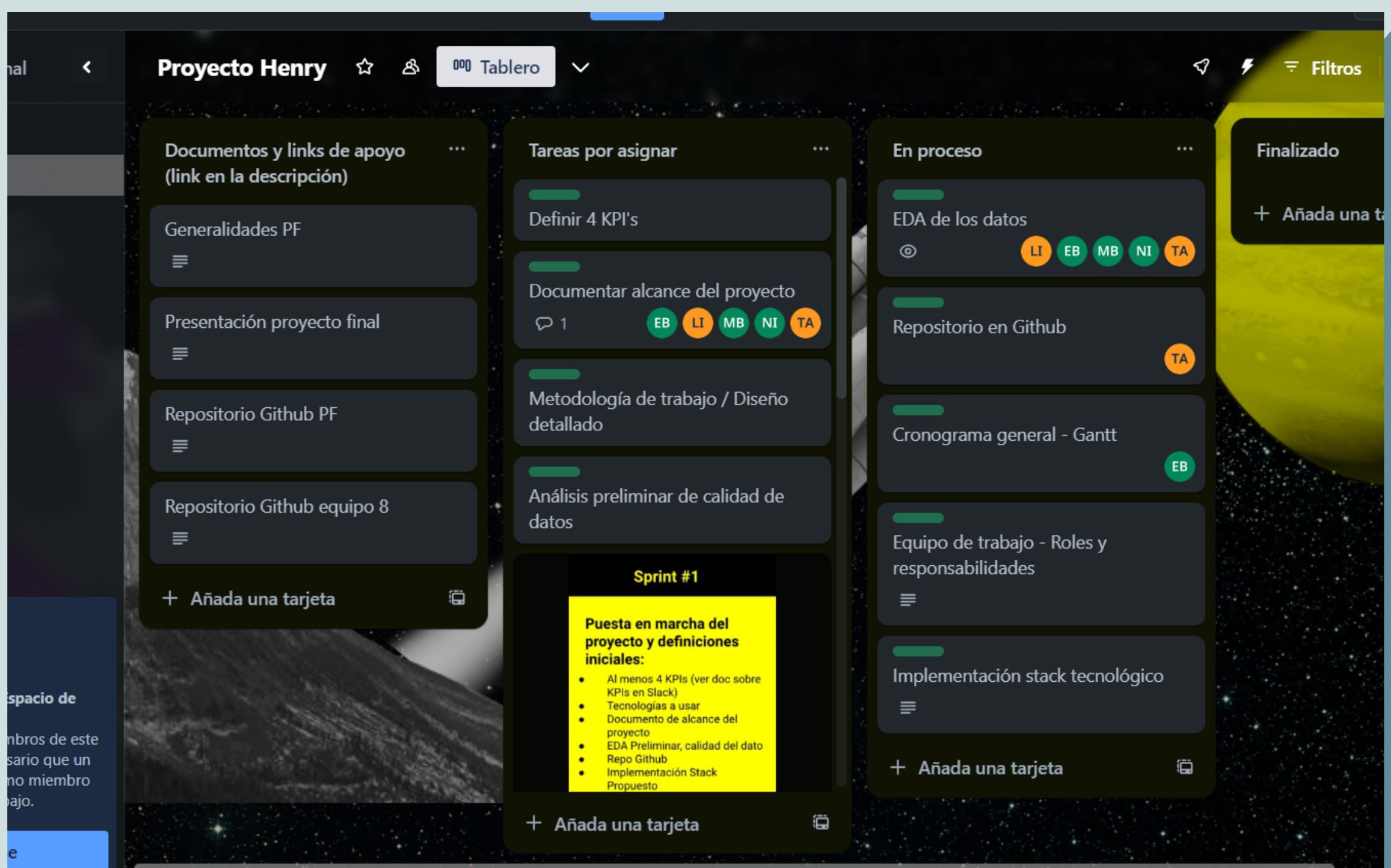


# Cronograma general

## Gantt



Implementamos esta herramienta gráfica con el propósito de visualizar de manera clara y precisa el tiempo estimado de dedicación para diversas tareas a lo largo de un periodo determinado. Este enfoque nos ha permitido coordinar eficientemente los plazos de entrega, optimizando la planificación del proyecto.



# Metodología de trabajo

El proyecto está planificado para una duración de 6 semanas, divididas en 3 sprints de 2 semanas cada uno.

A diario, los miembros del equipo trabajarán en las tareas asignadas a través de Trello. Al término de cada jornada, se llevará a cabo una reunión para compartir progresos y acordar nuevas tareas.

Además, se han programado 4 reuniones con el Henry mentor en cada sprint para discutir los avances y abordar posibles problemáticas que puedan surgir.

Al finalizar cada sprint, se llevará a cabo una presentación de los entregables durante una demostración que mostrará los avances logrados en esa etapa.

# EDA de los datos

The screenshot shows a Jupyter Notebook interface with the following content:

- Eliminamos columnas**:  
A code cell contains:

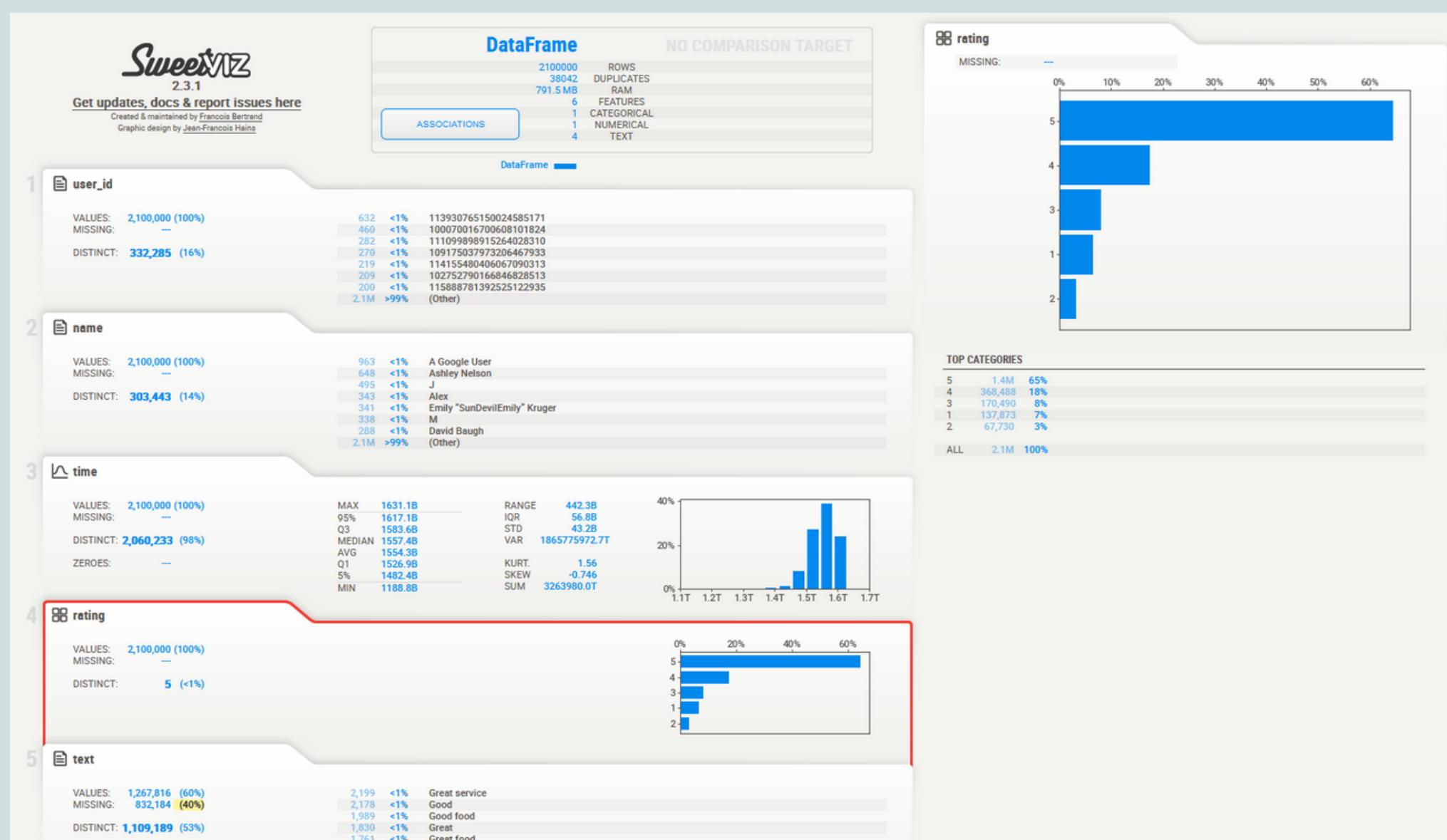
```
# Esta columna indica si un negocio esta abierto o cerrado en el momento, asi que no nos sirve.  
df.drop(columns='is_open', inplace=True)
```
- vemos si el id tiene duplicados**:  
A code cell contains:

```
# Contar la cantidad total de duplicados en la columna 'business_id'  
cantidad_duplicados = df['business_id'].duplicated().sum()  
  
# Mostrar la cantidad total de duplicados  
print("Cantidad total de duplicados en la columna 'business_id':", cantidad_duplicados)
```

Output: Cantidad total de duplicados en la columna 'business\_id': 0
- ¿Cuántos datos tenemos por Estado (state)?**:  
A code cell contains:

```
# Rellena los valores nulos con una cadena vacia  
df['state'].fillna('', inplace=True)  
  
conteo_por_estado = df['state'].value_counts()
```

Condujimos un análisis exploratorio inicial detallado, examinando la composición de los datasets. Evaluamos las columnas, las cantidades de datos, las distribuciones, así como la presencia de valores nulos y duplicados, entre otros. Este proceso nos brindó una comprensión integral de la calidad y la estructura de los datos.



# Alcances y objetivos del proyecto:

## Alcance temporal

3 Sprints de dos semanas cada uno.

## Alcance geográfico

Enfocado en los Estados Unidos, especialmente en aquellos estados con mayor crecimiento.

## Alcance de datos

Utilización de datasets de Yelp y Google Maps proporcionados por Henry, así como datasets externos del gobierno estadounidense sobre datos poblacionales.

## Identificar oportunidades para nuevos negocios

Utilizando un análisis exhaustivo del mercado, evaluando tendencias y preferencias, y considerando datos poblacionales para identificar áreas estratégicas de crecimiento.

## Analizar datos de usuarios para sistemas de recomendación

Aplicando técnicas de procesamiento de lenguaje natural y segmentación de usuarios para desarrollar sistemas de recomendación personalizados.

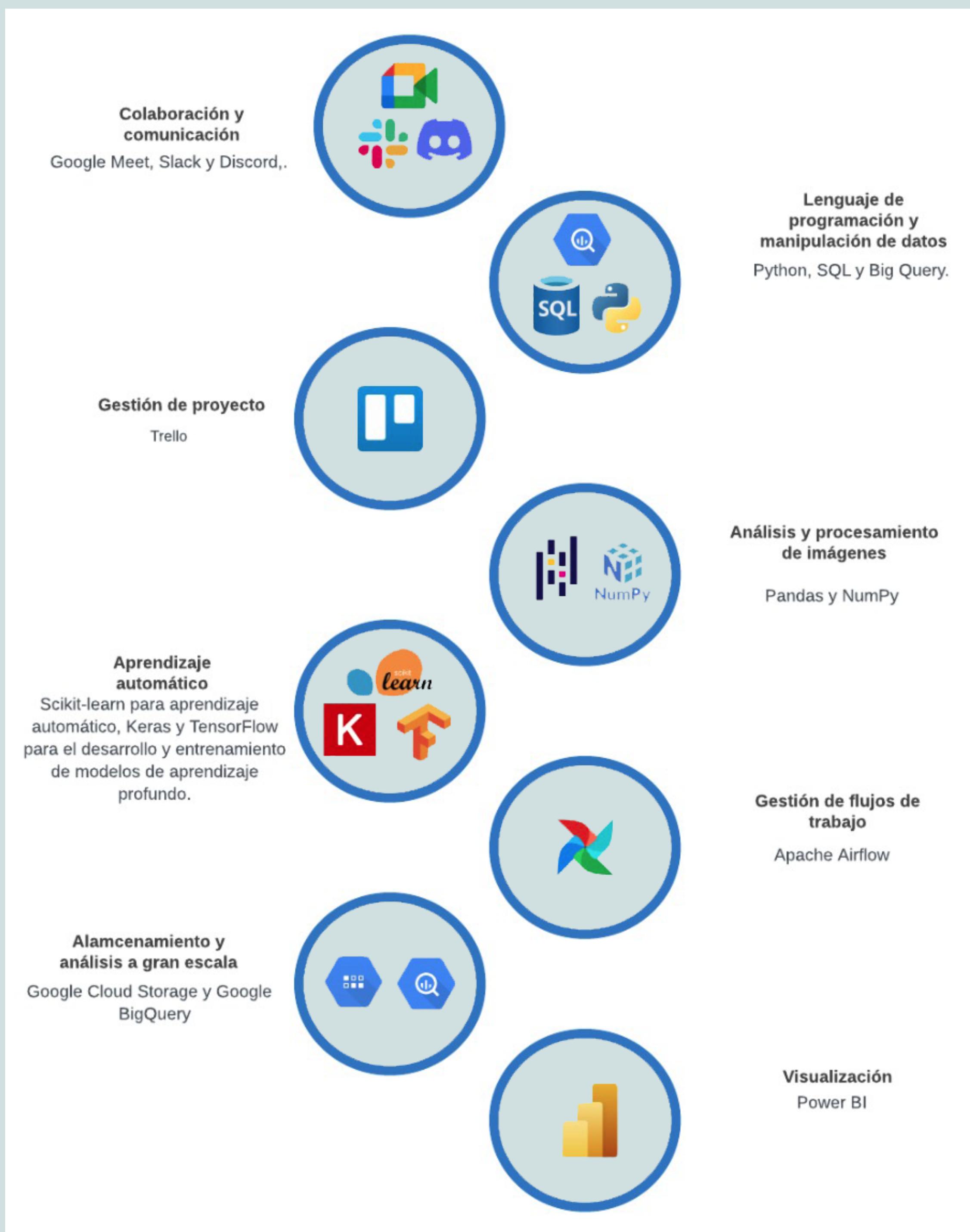
## Implementar modelos de aprendizaje automático

A través del preprocesamiento de datos, selección de variables relevantes y desarrollo de modelos predictivos evaluaremos el éxito potencial de nuevos negocios.

# KPI's

KPI	Descripción	Formula	Temporalidad	Objetivo
Variación trimestral rating	Seguimiento del progreso trimestral de las calificaciones de estrellas otorgadas por los clientes a cada local. Permite comprender las variaciones y tendencias a lo largo del año.	(promedio rating actual - promedio rating anterior) / promedio rating anterior ) * 100	Trimestral	5%
Variacion trimestral reseñas	Cálculo de la variación porcentual en el número total de reseñas recibidas por cada local de un trimestre a otro. Permite entender el cambio en la retroalimentación de los clientes.	(reseñas totales actual - reseñas totales anteriores / reseñas totales anteriores) * 100	Trimestral	5%
Reseñas positivas	Clasificación y análisis de las reseñas según su tono para cada local de una cadena, permitiendo identificar áreas de mejora y puntos fuertes específicos. La medida es la variación de un año a otro para evaluar su evolución.	[(reseñas positivas año actual - reseñas positivas año anterior) / (reseñas positivas año anterior)] * 100	Anual	10%
Reseñas Negativas	Clasificación y análisis de las reseñas según su tono para cada local de una cadena, permitiendo identificar áreas de mejora y puntos fuertes específicos. La medida es la variación de un año a otro para evaluar su evolución.	[(reseñas positivas año actual - reseñas positivas año anterior) / (reseñas positivas año anterior)] * 100	Anual	-10%
Presencia	Un ratio entre el numero de locales del nuestro rubro y el total de locales, esto permite ver la presencia y crecimiento del sector y la recepción del publico.	(numero de locales de nuestro rubro actual/numero de locales totales actuales) / (numero de locales de nuestro rubro año anterior/numero de locales totales año anterior) * 100	Anual	2%
Calificación vs competencia	Un ratio entre las calificaciones de nuestro rubro y el de la competencia permite observar el grado de satisfacción que ofrece el local vs la competencia. Al ponerle objetivo de tiempo podemos ver como evoluciona el grado de satisfacción que ofrece nuestros locales respecto a la competencia.	(calificación promedio nuestros locales actual/calificación promedio de la competencia actual) / (calificación promedio nuestros locales anterior/calificación promedio de la competencia anterior) * 100	Anual	2%

# Stack tecnológico



# Arquitectura



Comenzaremos con los datos almacenados en el drive, que incluyen archivos en formatos como JSON, CSV, Parquet y PKL. Nuestro enfoque será transferir toda esta información a Cloud Storage para llevar a cabo un proceso de limpieza utilizando Databricks. Posteriormente, migraremos los datos limpios a BigQuery, el cual utilizaremos como nuestro espacio de almacenamiento y análisis de datos a gran escala.

Después de esta fase, regresaremos a Databricks para realizar un proceso de ETL profundo y exhaustivo. Una vez completado, transferiremos los datos a BigQuery, dejándolos listos para su posterior análisis. Finalmente, utilizaremos Power BI para visualizar el informe.