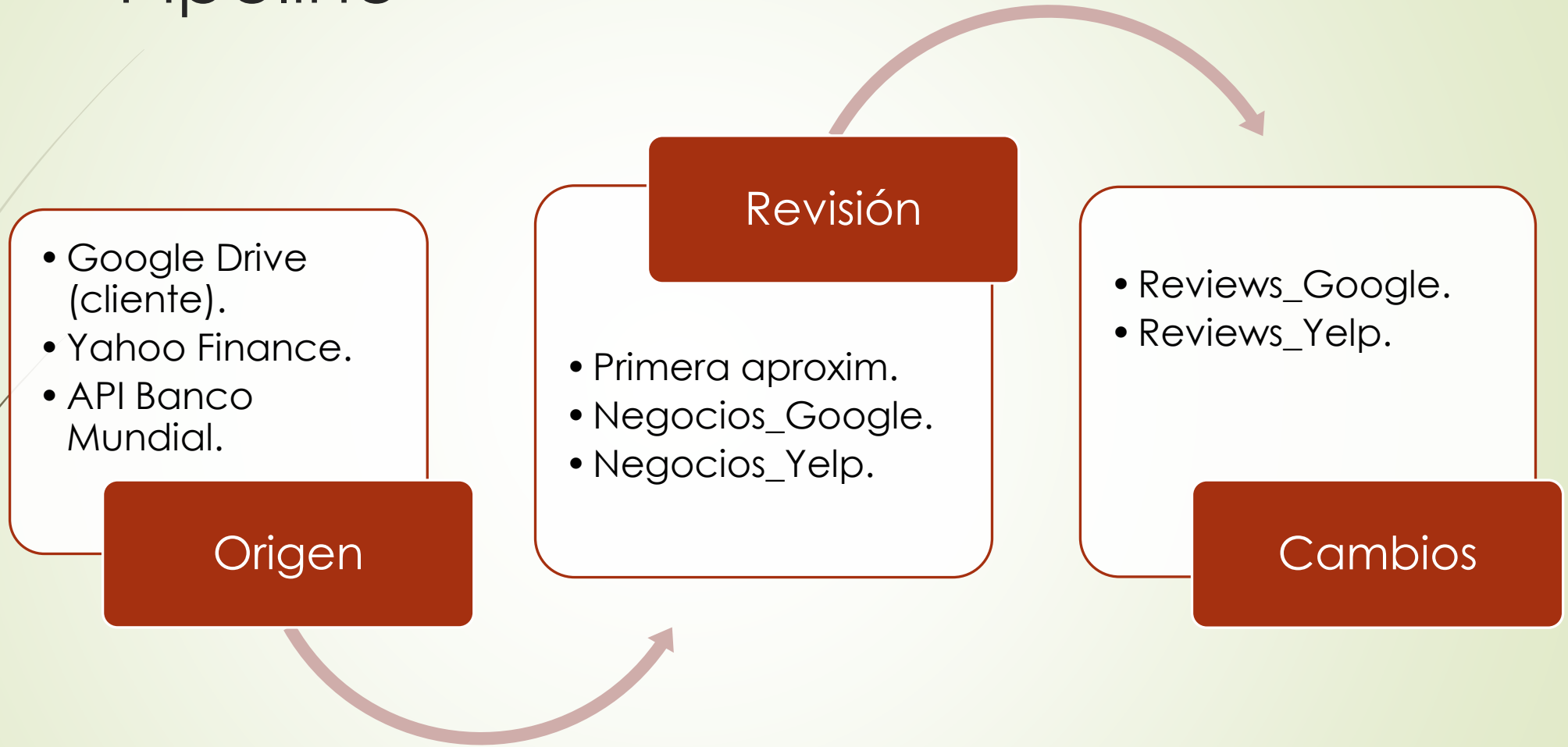
CIENCIA DE DATOS PARA RESTAURANTES

DATA ENGINEERING

Pipeline


• Google Drive (cliente).
• Yahoo Finance.
• API Banco Mundial.
Origen
Revisión
• Primera aproxim.
• Negocios_Google.
• Negocios_Yelp.
• Reviews_Google.
• Reviews_Yelp.
Cambios

Pipeline

Archivos seleccionados

Reseñas Google (51 estados) Aprox 20 mill

- Id usuario.
- Nombre restaurante.
- Fecha.
- Puntaje.
- Descripción.
- Id mapa.

Metadata sitios (11 archivos) 3,025,011

- Nombre restaurante.
- Dirección.
- Id mapa.
- Descripción.
- Latitud.
- Longitud.
- Categoria.
- Puntaje.
- Atributos.

Reseñas Yelp Aprox 7 mill

- Id reseña.
- Id usuario.
- Id negocio.
- Puntaje.
- Fecha.
- Reseña.

Business.pkl 150,346

- Id negocio.
- Nombre negocio.
- Ubicación.
- Rating.
- Numero reseñas.
- Esta cerrado.
- Atributos.
- Tipo comida.
- Horarios.

Pipeline

Archivos descartados

User.parquet

- Id usuario.
- Nombre usuario.
- Numero reseñas.
- Fecha creación.
- Id amigos.
- Votos por tipo.
- Fans.
- Años elite.
- Total cumplidos.

Checkin.json

- Id negocio.
- Fechas.

Tip.json

- Sugerencia.
- Fecha sugerencia.
- Total cumplidos.
- Id negocio.
- Id usuario.

Transformación

- ETL Google Restaurant.
  - Normalización de los campos.
  - Separación de la dirección.
  - Descarte de columnas.
  - Unión de DataFrames.
  - Valores nulos en Dirección.
  - Asignación de ID's.
  - Duplicados.
  - Total de registros: 204,702 vs 3,025,000 aprox.

Transformación

- ETL Yelp Restaurant.
  - Organización y descarte de columnas.
  - Asignación de ID's.
  - Duplicados.
  - Total de registros: 50,867 vs 150,000 aprox.

- Cruce Google Yelp.
  - Identificación de coincidencias (4,489 registros).
  - Homologación de Id's.
  - Combinación de DataFrames.
  - Total de registros: 251,080 vs 3,200,000 aprox.

Transformación

- Reseñas Yelp y Google.
  - Filtro de restaurantes.
  - Normalización de campos.

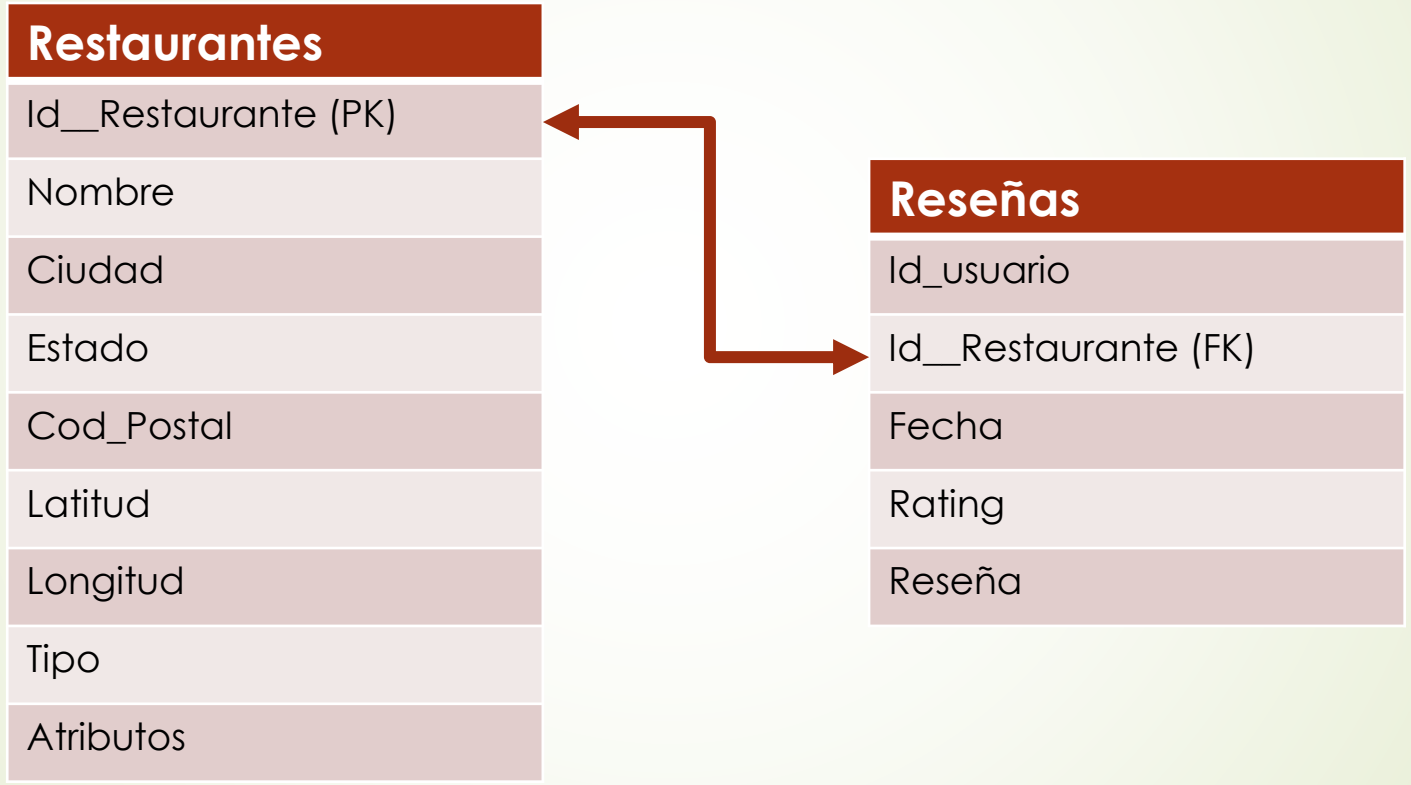
Stack Tecnológico

- Debido a la cantidad de información y, a los costos de la herramienta Azure, se decidió trabajar de manera local a través de Python, con sus librerías.
- Se utilizará Google Drive para el almacenamiento y accesibilidad de los archivos.
- Se analizará el uso de Azure para la implementación del modelo de Machine Learning o, dependiendo del rendimiento, valorar la posibilidad de trabajarlo de manera local.

Diagrama Entidad - Relación

- Tablas.
- Primary Key.
- Foreign Key.
- Diccionario de datos.

Tablas


Restaurantes
Id__Restaurante (PK)
Nombre
Ciudad
Estado
Cod_Postal
Latitud
Longitud
Tipo
Atributos
Reseñas
Id_usuario
Id__Restaurante (FK)
Fecha
Rating
Reseña

Próximas Tareas

- Normalización de atributos.
- Normalización de tipo de restaurante.
- Ordenar y homologar el repositorio Github.
- Armas tablas de Estado y Ciudad (Evaluar dependiendo tiempos y cargas de trabajo).