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.
• Normalización Estados.
Cambios

Pipeline

Archivos seleccionados

Reseñas Google (51 estados)

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

Metadata sitios (11 archivos)

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

Reseñas Yelp

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

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.

Busines.pkl

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

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

Transformación

- ETL Yelp Restaurant.
  - Organización y descarte de columnas.
  - Asignación de ID's.
  - Duplicados.
  - Total de registros: 50,867
- Cruce Google Yelp.
  - Identificación de coincidencias.
  - Homologación de Id's.
  - Combinación de DataFrames
  - Total de registros: 251,080

Stack Tecnológico

- (Herramientas y justificación de ellas).
- Debido a la cantidad de información, se decidió trabajar de manera local a través de Python, con sus librerías.
- Se utilizará Google Drive para el almacenamiento de los archivos.

Diagrama Entidad - Relación

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

Tablas

Restaurantes
Id__Restaurante
Nombre
Ciudad
Estado
Cod_Postal
Latitud
Longitud
Tipo
Atributos