LA ARQUITECTURA DE LOS DATOS

Los datos en la capa RAW Se encuentran en formato JSON

Los datos en las capas intermedias (stage) y los que ya estén listos para consumo, estarán en formato parquet, es decir trabajaremos con bases de datos columnares para aprovechar sus múltiples ventajas en entornos Big Data como por ejemplo eficiencia de almacenamiento y velocidad de acceso a datos. Esto es también posible debido a la estructura de las tablas proporcionadas que poseen pocas columnas, lo cual nos permite NO trabajar con un modelo relacional estricto, sin embargo se pueden reconcer alguna relaciones entre los datos proveidos.

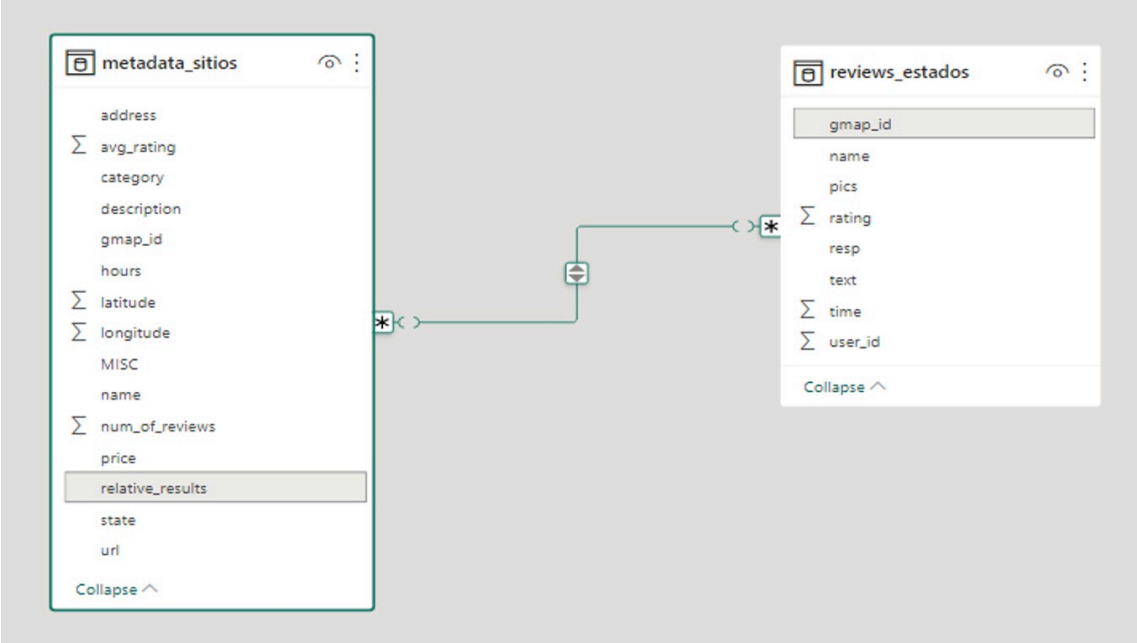

Fig.1 – Relaciones encontradas

Además los datasets están presentados con una configuración tipo “view”, como podemos apreciar en Fig.1, que podemos aprovechar con algunas transformaciones para su consumo posterior y llevarlo a una forma útil y consumible. Fig.2

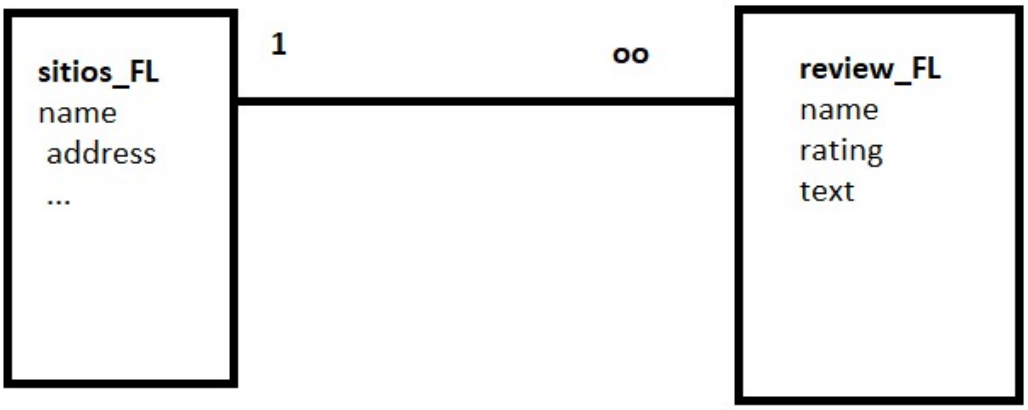

Fig.2 – Diagrama conceptual

Cargas de Datos (Data Loads)

Gestionaremos 2 tipos carga distintos debido a la naturaleza intrínseca de las tablas a usar:

Tabla dimensional: **Sitios_FL** -> Contiene información acerca de los sitios (negocios del rubro culinario). No poseen índices de tiempo. Para el mantenimiento / actualización se realizarán cargas totales con transformación posterior a la carga para la limpieza de datos.

Tabla referencial: **Florida_population_change** -> Contiene información demográfica del estado que será usada para proporcionar un marco contextual a los análisis realizados. Aquí también se empleará el tipo de carga total con transformación posterior a la misma para la limpieza de datos.

Tabla de Hechos: **Review_FL** -> Almacena información sobre registros de reviews otorgadas a los locales. Dado que esta información si tiene registro de tiempo se empleará cargas incrementales con un posterior tratamiento de limpieza de datos.

Así mismo existirán versiones especiales de algunas tablas que serán usadas especialmente en los modelos de machine learning.

El dashboard y el producto denominado ITakeU emplearan las mismas tablas.

DICCIONARIO DE DATOS

Tabla: Sitios_FL_ML.csv (Usada en Modelo de regresión supervisado)

Columna	**tipo**	**Descripción**
Latitud	Float	Latitud
Longitud	Float	Longitud
avg_rating	Float	Rating
price	Uint8	Categoría del precio
cat_bakery-desserts	Bool	Categoria del Negocio
cat_bakery-rest	Bool	Categoria del Negocio
Cat_restaurant	Bool	Categoria del Negocio

Tabla: birth_dataset (Empleada en Amazon Forecast)

Columna	**Tipo**	**Descripción**
Time	Datetime	Fecha
State	String	Sigla del Estado
Births	Int	Cantidad de nacimientos

Tabla: death_dataset (Empleada en Amazon Forecast)

Columna	**Tipo**	**Descripción**
Time	Datetime	Fecha
State	String	Sigla del Estado
Deaths	Int	Cantidad de muertes

Tabla: int_migration_dataset (Empleada en Amazon Forecast)

Columna	**Tipo**	**Descripción**
Time	Datetime	Fecha
State	String	Sigla del Estado
Int_migrations	Int	Cantidad de migraciones internacionales

Tabla: local_migration_dataset (Empleada en Amazon Forecast)

Columna	**Tipo**	**Descripción**
Time	Datetime	Fecha
State	String	Sigla del Estado
Local_migrations	Int	Cantidad de migraciones locales

Tabla: sitios_FL (Empleada en Quicksight y StreamLit)

Columna	**Tipo**	**Descripcion**
name	String	Nombre
address	String	Dirección
gmap_id	String	Id de Google maps
description	String	Descripción
latitude	Float	Latitud
longitude	Float	Longitud
category	String	Categoría del negocio
avg_rating	Float	Raiting promedio
num_of_reviews	int	Numero de reviews
price	String	Categoría de precio
MISC	Float	Misceláneos
state	String	Estado del negocio (activo o cerrado)
relative_results	String	Lista de Id de Google maps
url	String	url de mapa de local
street	String	Calle de la dirección
zip	int	Código postal
st	String	Estado de EEUU
City	String	Ciudad
horario	String	Horarios
open	String	Hora de apertura
close	String	Hora de Cierre
County Name	String	Nombre del condado

Tabla: city_population (Empleada en fase inicial de ETL para complementar transformaciones)

Columna	**Tipo**	**Descripción**
County_name	String	Nombre del condado
City	String	Ciudad
Poblacion	Float	Cantidad de habitantes

Tabla: reviews_FL(Empleada en Quicksight y StreamLit)

Columna	**Tipo**	**Descripcion**
name	String	Nombre del usuario
rating	float64	Rating otorgado
text	String	Texto de la review
date	Datetime	Fecha de review
zip	Int	Código postal