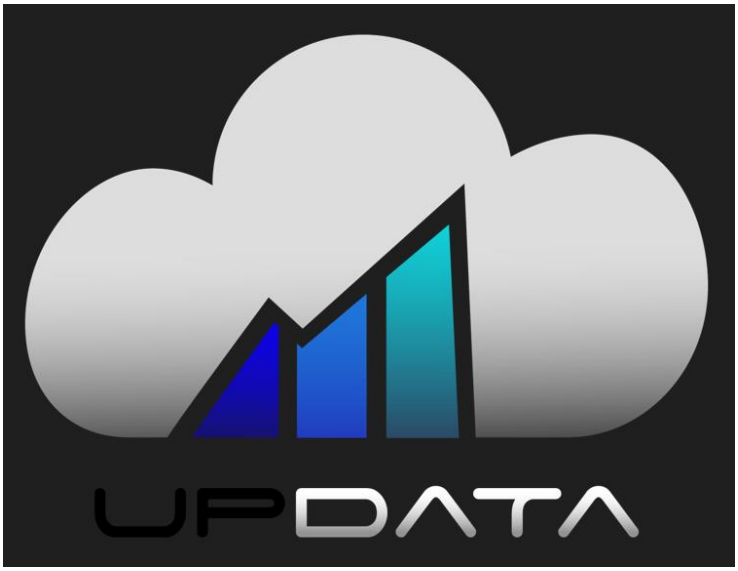
# PROJECT:

# PLACES TO EAT IN THE STATE OF FLORIDA, USA

**Data Consultant: UPDATA**




Ornaldo Hernández Ramos

Cristian Ignacio Papini

Aaron Bernardo Martínez Vera

Jhoeliel David Palma Salazar

Janice Rocío Rico Sánchez

# INDEX

# INTRODUCTION

Being in the Age of Data and where the most important are those that come directly from people, the reviews that are made about places to eat throughout the United States throw up important points to be considered. These are useful to know the preferences of potential customers, as well as valuable information when opening new stores.

That is why an analysis of this US market must be carried out, to rely on the opinion of those who have the experience of going to these places and tasting what they offer, serving as a knowledge path for those who are then willing to do the same.

The power of customer reviews leaves a range of possibilities when making decisions not only for other users, but also for businessmen interested in this business area.

Google Maps offers a large amount of information about restaurants that mainly details their location, attributes and categories. Likewise, the opinions of the users about these places are collected, which marks a usable thread to be taken into account in later solutions.

# OBJECTIVES

## General:

- ✓ Provide accurate and timely information on places to eat in the State of Florida, to diners and businessmen, so that they make better decisions, each one in their area.

## Specifics:

- ✓ Build an efficient and comprehensive Data Warehouse, with reliable data, using both the Datasets provided by the client, as well as others acquired externally by the Data Engineering team.
- ✓ Analyze the opinion and rating of users in Google Maps regarding their interaction with places to eat in the State.
- ✓ Analyze the restaurant locations with the best ratings.
- ✓ Give recommendations to diners and businessmen, based on the results of previous studies.
- ✓ Present an interactive web tool for users that allows them to make different inquiries about places to eat.
- ✓ Conform Dashboards with data analysis graphs, so that entrepreneurs identify business opportunities in the field of restaurants.
- ✓ Extract and present KPIs to evaluate the performance of information-generating tools.
- ✓ Deliver a Machine Learning report with suitable Models that show predictions that help decision making for future investments.

# SCOPE

Through this Project, EDA and ETL processes will be applied to the data supplied from Google Maps user sites and reviews. In addition, they will be complemented with others obtained from various sources, thus achieving an interactive platform for customers to consult options for places to eat in the State of Florida, according to key data entered.

In the same way, informative dashboards will be generated for entrepreneurs interested in investing in the category of food places, with data crossing that will result in understandable and valuable figures and graphs for decision making.

Why focus on Florida?

- According to the United States Census Bureau, Florida has a population of 21,538,187 inhabitants (Census of April 1, 2020), of which more than 87% have a computer and a Internet subscription, and 591,046 establishments that are a source of employment. Favorable data for a good choice of places to eat and invest.

(https://www.census.gov/quickfacts/fact/table/FL#)

- According to the United States Bureau of Economic Analysis (The U.S. Bureau of Economic Analysis - BEA), the Percentage Change in Personal Consumption Expenditures is included in the range of the highest, from 10.4% to 12.5%.

(https://www.bea.gov/data/consumer-spending/state)

- According to the National Restaurant Association, there are about 40,000 eating and drinking establishments in Florida, generating $42 billion in revenue. Some of the most famous cities are Orlando, Miami, Tampa, West Palm Beach and Boca Raton.

(https://restaurant.org/) (https://www.restohub.org/operations/planning/florida-restaurant/)

- On June 22, 2022, in an article published on the Forbes website, Gilda D'Incerti, Founder and CEO of PQE Group, a Technology Solutions and Consulting Services Company, and also a Forbes Board Member, spoke of “3 Reasons Why My Company is Moving to Florida”. It is the 4th. State with the highest economic growth (around 30% between 2011 and 2021).

In 2019, Forbes ranked Florida 5th. spot on its Best States to Do Business List, detailing its high average rate of net migration and projecting that its job and income growth would outpace the nation over the next five years. Florida also ranks first in the Kauffman Early-Stage Entrepreneurship Index, with very promising metrics such as 86% of new entrepreneurs starting a business there out of choice rather than necessity, and a start-up survival rate of 80%.

**1st. Reason:** Great Growth in Jobs and Wages. According to the US Census, Florida's population growth over the past decade was 14.6%, the highest of any state. Florida has also recently seen an influx of remote workers due to its low cost of living, and the Northeast particularly seems to be a popular area for it.

It's no wonder Florida is so attractive to workers, since in addition to enjoying a low cost of living, they earn more. Wages in the state increased 9.7% from December 2020 to December 2021, the second largest increase in the US As more companies open offices in Florida, the job market in the State is also booming. The unemployment rate dropped from 8.9% to 3% between April 2021 and April 2022.

**2nd. Reason:** An Excellent Location for Business. Florida ranks so high as a good place for business because it includes business-friendly policies and its regulatory environment is simplified. The State offers incentives to attract all types of companies.

There are also 16 international passenger service airports there, and several, such as Tampa, Orlando and Miami, offer direct flights to Europe. In addition, there are several new transportation developments in the pipeline, including rail systems.

Florida's tax policies are another draw. It is one of nine states with no income tax, and there is no tax on inheritances, gifts, or intangible personal property such as stocks. Additionally, corporations have the benefit of only paying 5.5% on their tax return and no payroll taxes. That can be a game changer for a business just getting started or an established business looking to leave a high-tax state.

**3rd. Reason:** The Sunshine State. It's the warmest state in the US, with lots of sunshine and an average annual temperature of 70.7°F. The southernmost state in the US; its northern and central areas have a subtropical climate, and its southern areas are tropical. Its excellent climate throughout the year means that almost every day is a beach day. (https://www.forbes.com/sites/forbesbusinesscouncil/2022/06/22/three-reasons-why-my-company-is-moving-to-florida/?sh=72fbe8175e9d)

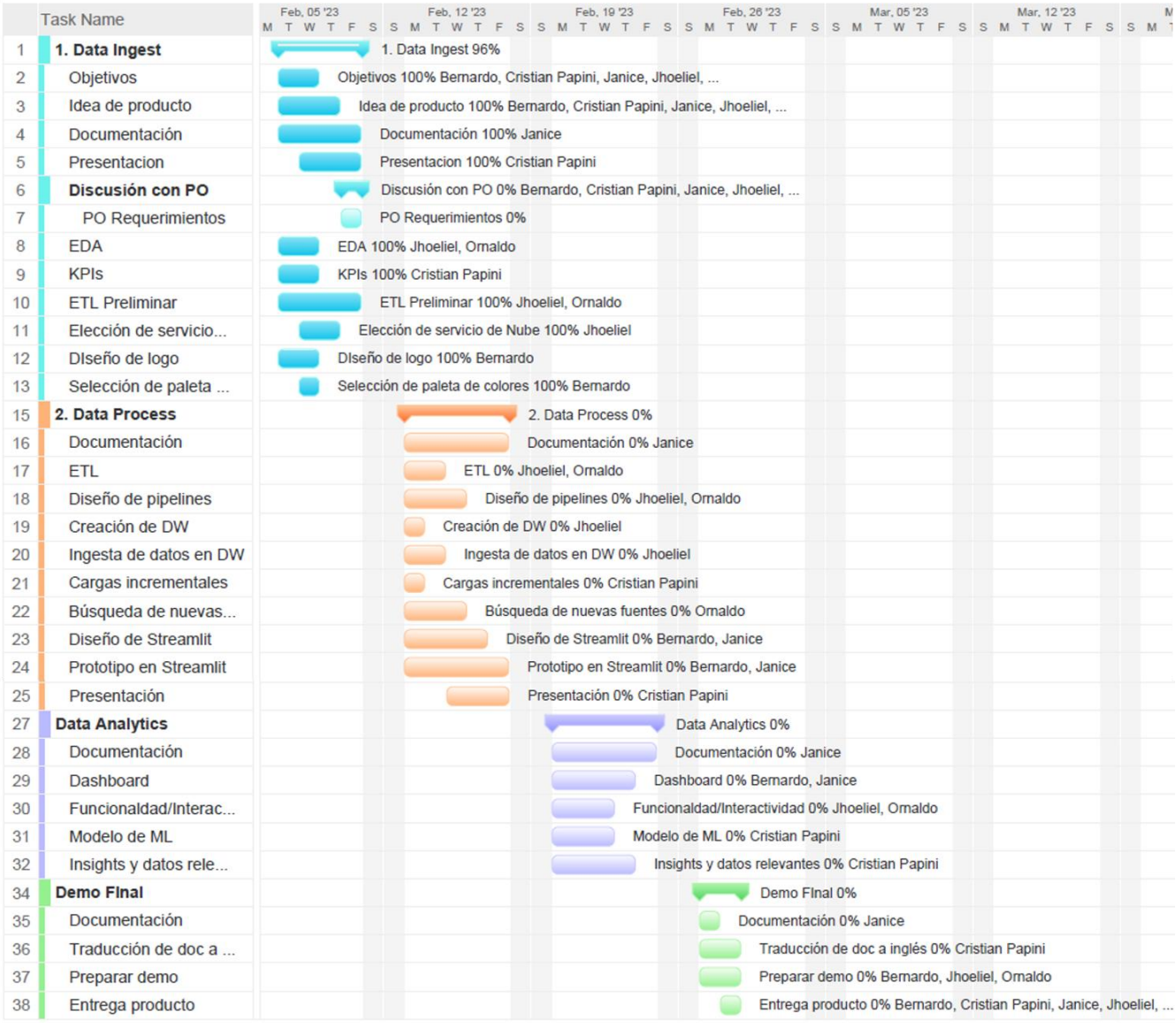
# WORK METHODOLOGY

For the execution of this Project, from the first steps of analysis to the final product, an Agile Methodology of the Scrum type was carried out, with dailies of the work team where the progress of the previous day and the tasks for the current one were discussed.

The progress and goals achieved were reported to the Product Owner on a weekly basis, as well as receiving his feedback.

Goals were established for a month (general), for each week, and daily. For this, a Gantt Diagram was prepared, with the role of each team member and the estimated times for its execution.

# GENERAL SCHEDULE - GANTT CHART



# WORK TEAM – ROLES AND RESPONSIBILITIES

The different members of the team had different roles according to their abilities and personal decision to work in a certain area, which increased the efficiency of the tasks that were carried out during the execution of the same. This is how each one reported their doubts and results to the other members, so that the objectives set and the required deadlines were met.

The work was distributed as follows:

Data Engineer: Ornaldo Hernández.

Data Cloud Engineer: Jhoeliel Palma.

Data Analytics: Aaron Martínez y Janice Rico.

Machine Learning Engineer: Cristian Papini.

# DETAILED DESIGN - DELIVERABLES

Each of the work files and visualizations were stored in a GitHub Repository (https://github.com/janicerico/PG_Google_Maps.git), which contains:

- ✓ Jupyter Notebooks for the EDA processes, and automated ETL for the loading and treatment of the initial Datasets.
- ✓ Final datasets from which the WebApp in Streamlit, the Dashboards and analysis with their KPIs, and the application of Machine Learning Models were generated.
- ✓ Interactive platform to make inquiries regarding Recommended Places to Eat for diners, as well as extra Statistics for future Investors.
- ✓ Dashboards with AWS tools, including the KPIs generated, for entrepreneurs.
- ✓ Visualizations of the results of Machine Learning Models.

# INITIAL DATA PROVIDED

The supplied Datasets are divided into several folders with the following distribution:

| File | Sub-File | Quantity .json files | Weight (Mb) |
|---|---|---|---|
| reviews-estados | review-Alabama | 12 | 467,90 |
| | review-Alaska | 4 | 142,30 |
| | review-Arizona | 14 | 636,00 |
| | review-Arkansas | 16 | 611,80 |
| | review-California | 18 | 762,90 |
| | review-Colorado | 16 | 751,50 |
| | review-Connecticut | 18 | 673,60 |
| | review-Delaware | 7 | 235,00 |
| | review-District_of_Columbia | 4 | 160,90 |
| | review-Florida | 19 | 886,30 |
| | review-Georgia | 13 | 575,50 |
| | review-Hawaii | 11 | 471,70 |
| | review-Idaho | 14 | 569,50 |
| | review-Illinois | 14 | 587,50 |
| | review-Indiana | 15 | 599,00 |
| | review-Iowa | 18 | 673,00 |
| | review-Kansas | 13 | 511,00 |
| | review-Kentucky | 11 | 429,50 |
| | review-Louisiana | 10 | 385,30 |
| | review-Maine | 8 | 290,00 |
| | review-Maryland | 16 | 669,30 |

| | | | |
|---|---|---|---|
| | review-Massachusetts | 16 | 640,60 |
| | review-Michigan | 15 | 624,70 |
| | review-Minnesota | 12 | 502,80 |
| | review-Mississippi | 14 | 486,70 |
| | review-Missouri | 11 | 455,90 |
| | review-Montana | 7 | 248,00 |
| | review-Nebraska | 13 | 468,30 |
| | review-Nevada | 12 | 520,00 |
| | review-New_Hampshire | 9 | 342,40 |
| | review-New_Jersey | 13 | 539,80 |
| | review-New_Mexico | 12 | 486,90 |
| | review-New_York | 18 | 759,90 |
| | review-North_Carolina | 15 | 656,20 |
| | review-North_Dakota | 4 | 144,90 |
| | review-Ohio | 13 | 538,50 |
| | review-Oklahoma | 11 | 445,80 |
| | review-Oregon | 15 | 644,90 |
| | review-Pennsylvania | 16 | 658,30 |
| | review-Rhode_Island | 6 | 227,20 |
| | review-South_Carolina | 14 | 575,60 |
| | review-South_Dakota | 5 | 172,60 |
| | review-Tennessee | 12 | 497,40 |
| | review-Texas | 16 | 713,50 |
| | review-Utah | 10 | 477,20 |
| | review-Vermont | 3 | 85,60 |

| | | | |
|---|---|---|---|
| | review-Virginia | 12 | 480,80 |
| | review-Washington | 13 | 569,80 |
| | review-West_Virginia | 8 | 264,80 |
| | review-Wisconsin | 12 | 454,10 |
| | review-Wyoming | 3 | 111,00 |
| | | **Total weight (Mb)** | **24883,70** |
| | | | |
| metadata-sitios | | 11 | **2832,30** |

# DATA COLLECTION AND TRANSFORMATION PROCESS

From all the data and the business decision for places to eat in the State of Florida, two sets of data were selected: The group of 11 .json files of "metadata-sites" that contained information of establishments in general sense (regarding usefulness and geographic belonging), and another group of 19 .json files from “review-Florida” containing reviews about businesses.

The information about the businesses (“metadata-sites”) received different transformations:

1. Removed '\n' from character columns.

2. After parsing the table, changes were made to the data types of some columns to another data type that took up less space and could store the same information.

3. Removed the string-to-string list data type in two columns.

4. The 11 .json files were saved in only one with a .parquet format, and another in .csv.

5. Duplicates were removed.

6. The data was filtered by different criteria: RESTAURANT, BAKERY, DESSERTS, PASTRY, and from Florida.

7. New columns were created, starting from the address column, for the Analytics and Machine Learning work as: st (state) and zip (zip code). Other columns were created to mark exactly what type of establishment it was: Restaurant, Bakery, Desserts, or Pastry.

8. Two columns were created: open and close, to indicate opening and closing hours of the establishment; only when possible.

9. Two new columns could be created from the zip column: county and city. Taken from http://census.gov.

10. An attempt was made to obtain, via scraping, the geographic coordinates of each of the establishments to improve those that came in the initial data, from the address column. The found method worked fine, but it was based on the Selenium module that allows you to simulate a browser to render the html generated by the url, and to get the coordinates it took a long time. This had to be done for more than 4,000 establishments, and after three days almost full time, it barely reached 1,000. Then it was decided to continue with the coordinates that came originally.

Starting from the “review_FL” table (unified table of all reviews – 19 .json files):

1. Two new columns were created: avg_rating (average ratings for an establishment) and num_of_reviews (number of reviews).
2. Added five columns to it: rating1, rating2, rating3, rating4, and rating5, which have the sum of the ratings that match their name rating1 = sum of ratings = 1, and so on.

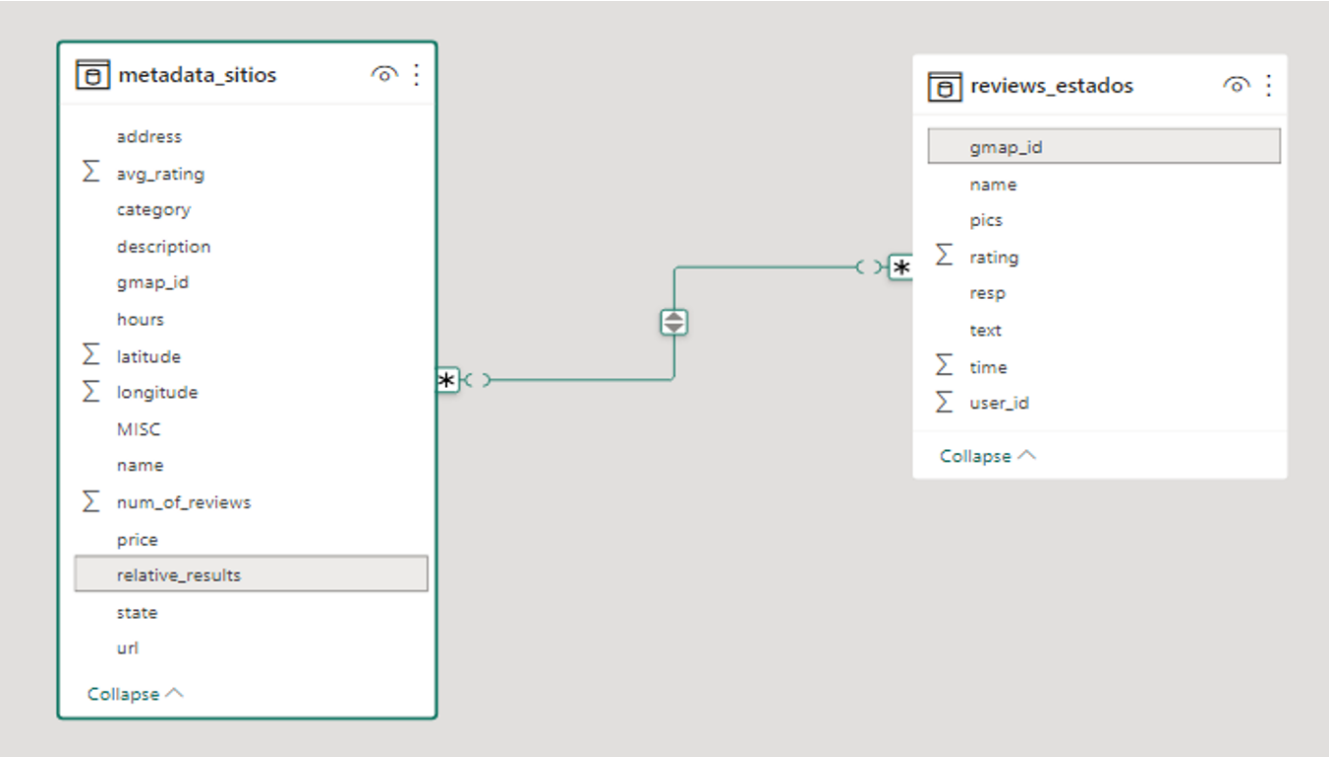
Additionally, extra information was obtained for the Analytics team and the Machine Learning team to decide if it had value or not:

- ✓ **city_population**: Population of each city in Florida.
- ✓ **Florida_population**: Population balance of Florida, between 1998 and 2020.
- ✓ **GDP**: Gross Domestic Product of each State of the United States, between the years 2021 and 2022. (http://census.gov)

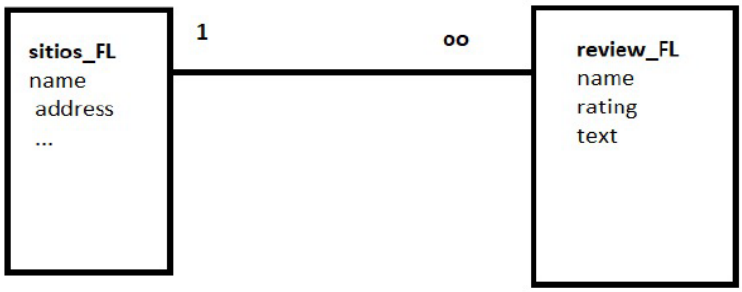
# DATA ARCHITECTURE

The data in the RAW layer is in .json format.

The data in the intermediate layers (stage) and those that were ready for consumption are in .parquet format. Columnar databases were used to take advantage of their multiple advantages in Big Data environments, such as storage efficiency and speed of data access. This is also possible due to the structure of the provided tables that have few columns, which allowed **not** to work with a strict relational model. However, some relationships could be recognized between the data provided:




In addition, the Datasets are presented with a "view" type configuration, which could be used with some transformations for later consumption and take it to a useful and consumable form.



# DATA UPLOADS TO AWS CLOUD

Two different load types were managed due to the intrinsic nature of the tables to be used:

Dimensional table: sitios_Fl -> Contains information about the sites (businesses in the culinary industry). It does not have time indices. For maintenance / updating, total loads were carried out with subsequent transformation for data cleaning.

Reference table: florida_population_change -> Contains demographic information from the State that was used to provide a contextual framework for the analyzes performed. Here, too, the full load type with post-transformation was used for data cleansing.

Fact Table: review_Fl -> Stores information about records of reviews given to venues. Since this information does have a time record, incremental loads were used with a subsequent data cleaning treatment.

Likewise, there were special versions of some tables that were used especially in the Machine Learning, Dashboard and WebApp iTakeU models.

# DATA DICTIONARY

Table: sitios_Fl_ML.csv (Used in Supervised Regression Model)

| 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 |

Table: birth_dataset (Used in Amazon Forecast)

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

Table: death_dataset (Used in Amazon Forecast)

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

Table: int_migration_dataset (Used in Amazon Forecast)

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

Table: local_migration_dataset (Used in Amazon Forecast)

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

Table: sitios_Fl (Used in 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 |

Table: city_population (Used in the initial phase of ETL to complement transformations)

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

Table: reviews_Fl (Used in 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 |

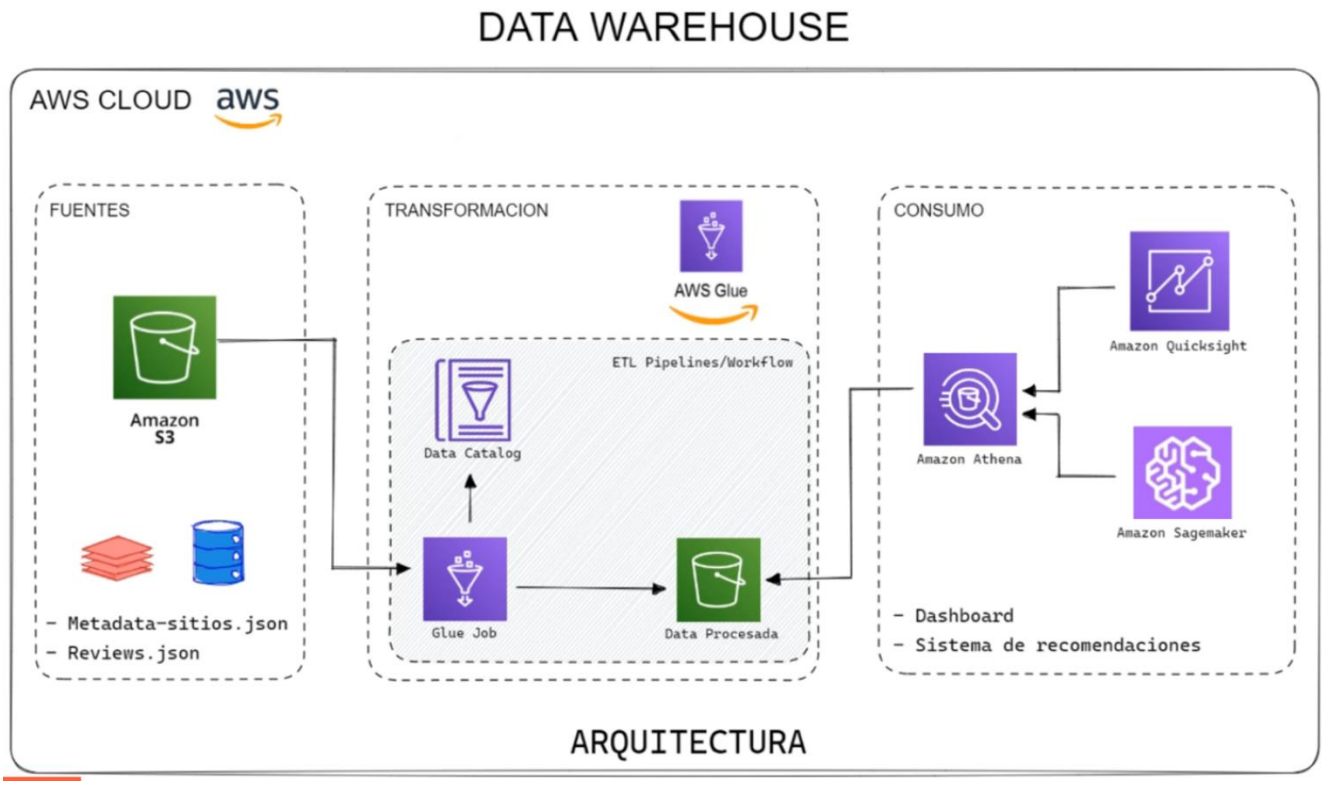
# ETL PROCESS IN AMAZON CLOUD

Explanatory Video ETL Process in Amazon AWS:

https://www.youtube.com/watch?v=irX-6nPQTjI

Video Showing the Process of Implementing the ETL Process in AWS Cloud - Quick View:

https://www.youtube.com/watch?v=3-nzZq9ZcF8



# DASHBOARD

At the beginning of the development of this project, the use of professional tools such as AWS Cloud and its services was proposed to facilitate the transformation of information.

First, an investigation of the operation of the extension to be used (Quicksight) was generated. Then, basic information was sought and dedicated to its correct handling.

Therefore, the Dataset inside the Buckets was obtained and studied (familiarized with the different columns and the information they displayed). Finally, the most relevant fields were taken to graph, depending on the information to be displayed and the KPIs developed in the other phases.

The tool that was used was "Quicksight", which facilitates the visualization of the data in graphic form and in a format that consists of selecting and dragging the desired fields, and the visual elements were adjusted with a user-friendly color palette.

# KPIs

In order to make decisions and keep track of the behavior of the different variables in the project, KPIs were developed, which are metrics evaluated over time that emphasize variations.

Below are the different KPIs proposed and the questions and/or ideas to which they respond:

**Count Reviews by Category:**

Information: Number of reviews by category.

Questions to which it answers:

In which category will I have more information on which to support my decision? More information = More decision-making capacity.

In which category do I have the most support/follow-up from customers? Greater contact with the client = Greater ability to establish what the client likes or dislikes.

**Rating by Category:**

Information: Average rating by category.

Questions to which it answers:

Which category has the best reputation?

What category is the most acclaimed by people? I would not want to bet on categories little recognized by the public.

**Count Reviews by Day:**

Information: Number of reviews by day of the week.

Questions to which it answers:

What days of the week should I open the place?

Is it convenient to have open 7 days a week? Beyond the fact that this is defined through an economic-financial analysis, with the number of reviews you can get an idea of what the flow is like throughout the week.

**Rating by Price:**

Information: Average rating by price type.

Question to which it answers:

What type of market is the most acclaimed by people? Inexpensive, intermediate or high-end?

This helps me, beyond the ideas I may have about the business, to adapt my idea to the current market demand.

# MACHINE LEARNING MODELS

Taking into account the Datasets that were available and the business idea that was intended, different Machine Learning models were devised that would serve as tools for the client to make decisions in the future.

## Prediction Model – Amazon Forecast

For this, a Prediction Model was first created with the Amazon Forecast tool, in which the Datasets of Births, Deaths, International and Local Migrations were uploaded, with information from 1990 to 2020. After creating predictors in the tool according to the data available (in this case, the tool chooses the predictor that best suits the Dataset), a forecast was trained and created. The latter had the ability to predict up to 7 years after the last date available, in this case 2020, therefore, the forecast reached the year 2027.

As a result, for each of the 4 variables mentioned, a series of data was obtained with which a graph and its respective percentiles were constructed. These helped in understanding and reading the forecast.

If the population behavior of the State is added to the analysis of the GDP (Growth Domestic Product) and the PCE (Personal Consumption Expenditures), growth and consumption respectively, it would be in a position to predict if the State of Florida is a good option for investors.

## Supervised Classification Model – Amazon Sagemaker

After analyzing the aforementioned forecasts, it was decided that the State of Florida is a good place to make gastronomic investments. It is known that this category is too broad and that is why it was decided to develop a Machine Learning model that predicted, given certain conditions, the estimated Rating of the place.

For this, the Amazon Sagemaker tool was used to develop a Machine Learning Model in which Random Forests were applied with an accuracy of 68%.

The inputs that the user of the model must enter are:

- Coordinates of the place.

- Price type (0,1,2,3).

- Category.

Based on these 3 inputs, the model is able to predict the estimated rating that the place should have.