Sprint #2

Data Engineering

Arranca con extraer los datos de la api, se hace el etl (desde glue, se automatiza con ariflow), y despues los cargamos a la base de datos em mysql, despues el primer paso lo vamos a dejar automatizado en un pipeline cargado por airflow en un linux que va a estar montado hasta ahora en EC2, pero nose si van a usar otra plataforma al final

El módulo **pyspark.ml.recommendation** de PySpark proporciona una clase llamada ALS (Alternating Least Squares) que se utiliza para crear modelos de recomendación.

CLOUDibm2023+2023

bussines_google.json.gz dataset google

bussiness_yelp.json.gz dataset yelp

amazon s3 data warehouse
rds amazon gestor de bases de datos

Hitos

- ETL completo
- Estructura de datos implementada (DW, DL, etc). Pueden usar algún servicio
- Pipeline ETL automatizado
- Diseño del Modelo ER
- Pipelines para alimentar el DW
- Data Warehouse
- Automatización
- Validación de datos
- Documentación
    - Diagrama ER detallado (tablas, PK, FK y tipo de dato)
    - Diccionario de datos
    - Workflow detallando tecnologías
- Análisis de datos de muestra
- MVP/ Proof of Concept de producto de ML ó MVP/ Proof of Concept de Dashboard

ETL completo

Seguimiento de actividades fundamentales durante el proceso ETL de los datos

- Carga incremental para unificar datasets: “Estados” - “Metadata google”
- Aparente incoherencia con datos de coordenadas con relación a su estado perteneciente

Eliminar columnas redundantes
- no son consideradas relevantes para el análisis
- se pueden obtener utilizando una URL
- cantidad de datos escasa y poco relevante

filtro por estados de interés (New jersey, California, Florida, Illinois)
filtro por categorías de restaurantes
filtro por las reviews del 2015 en adelante
Exportación de los archivos .gzip - Google drive, Github

- Carga incremental para concatenar archivos .json: creación de nuevo archivo “Reviews”

- Se normaliza la columna 'time' convirtiendo valores en formato de fecha y hora
- Seleccionar los estados de interés, cargar y concatenar los datos de los restaurantes por categoría
- Análisis de sentimiento

Exportación de archivo .parquet - Google drive, Github

Área Staging

Minimizar errores
Transformación de datos
Optimización del mapeo de datos
Independencia de la planificación
Soporte para archivo de datos
y
Resolución de problemas

pandas

NLTK

APACHE ARROW

NumPy

PySpark

seaborn

docker

Estructura de datos implementada

2. Este Hito implica la implementación de la estructura de datos en un Data Warehouse (DW), Data Lake (DL), entre otros. Esto puede implicar la creación de esquemas de bases de datos, la definición de relaciones entre las tablas, la implementación de índices para mejorar el rendimiento de las consultas, entre otras tareas.

RDS , EC2, GLUE
la pc virtual de ec2 para el pipeline automatizado de etl con glue,

Data Warehouse

Despliegue más rápido
Administración eficiente de los datos
Durabilidad y disponibilidad
Seguridad

Pipeline ETL Automatizado

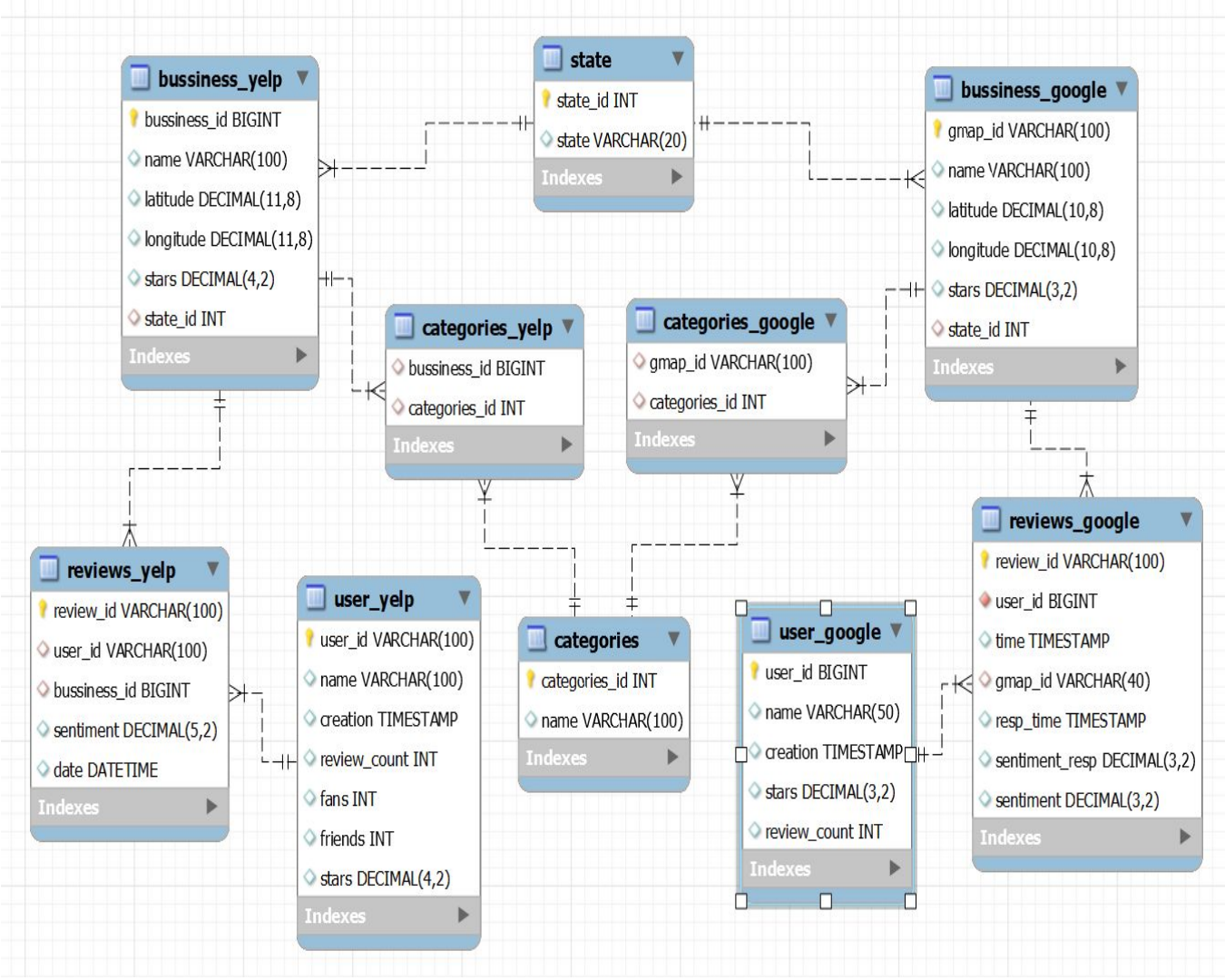
Diseño del Modelo ER (Entidad - Relación)

Representa las entidades en nuestro sistema y cómo se relacionan entre ellas, permitiendo la visualización de las tablas de datos y sus relaciones.

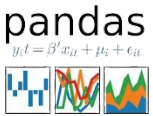
facilitando la comprensión de cómo se almacenan, su manipulación y extracción, su relación e identificación de elementos clave.

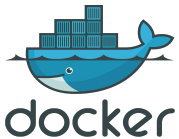
MySQL Workbench

“ El ID en un modelo ER es crucial para la gestión eficiente de datos, ya que proporciona una identificación única para cada entidad, facilitando la organización, búsqueda y manipulación de los datos.”



Workflow detallando tecnologías





E-T-L
Extracción-Transformación-Carga

Para montar la infraestructura de tu proyecto con pipelines para realizar el proceso de ETL apuntando a estructuras de tipo Data Warehouse, Data Lake o Datalakehouse, y contemplando la carga incremental de datos, puedes seguir los siguientes pasos:

1. Identificar los datos de origen: El primer paso es identificar los datos de origen que se van a procesar. Estos datos pueden provenir de diversas fuentes como bases de datos de producción, CRM, APIs, entre otros.

2. Diseñar el área de staging: Una vez identificados los datos de origen, el siguiente paso es diseñar el área de staging. Esta es la etapa donde los datos de origen se transforman y se preparan para ser cargados en el Data Warehouse.

3. Definir el esquema del Data Warehouse: El esquema del Data Warehouse debe ser definido de acuerdo a las necesidades del proyecto. Este esquema determinará cómo se organizarán y almacenarán los datos en el Data Warehouse.
4. Implementar la lógica de carga de datos: La lógica de carga de datos se encarga de mover los datos desde el área de staging hasta el Data Warehouse. Este proceso puede ser incremental, lo que significa que solo se cargarán los datos nuevos o modificados desde la última carga.
5. Cargar los datos de forma incremental: Para cargar los datos de forma incremental, puedes utilizar técnicas como Change Data Capture (CDC). Esta técnica captura todas las operaciones mutantes en un sistema de base de datos de origen y las transmite a otro sistema. CDC mantiene todos los cambios intermedios, incluyendo los eliminados. Con esta arquitectura, no solo capturas las inserciones y actualizaciones, sino también los eliminados confirmados en el Data Lake, y luego fusionas esos cambios capturados en los Data Warehouses.

6. Testear y monitorizar la pipeline ETL: Una vez implementada la lógica de carga de datos, es importante testear y monitorizar la pipeline ETL para asegurar que los datos se están cargando correctamente y de manera eficiente **(Para implementar un pipeline ETL automatizado, puedes utilizar diversas herramientas y tecnologías como Apache Airflow, AWS Glue, Azure Data Factory, entre otras. Estas herramientas te permiten diseñar, programar y monitorear tus pipelines ETL)**

ETL - reception Bruno

estados, reviews archivos json, nomralizar columna time resp
estados de interes, cargar, concatenarlos, bussiness(restaurant de categoria)
fechas reviews, 2016 - adelante
respuestas a reviews, analis de sentimiento
nltk, no optimo
-1 , 1 ponderacion con rating de lkocal dada por usuario
sentiment muy buneo,bueno, malo,neutral
reviews limpio
subir parquet

Problemática coordenadas

** agregar comentario, problemática y solución, jona **

Identificamos discrepancias en la asignación de estados en la columna "estado" de nuestro conjunto de datos de empresas de Google Maps. La información de la columna "estado" no coincidía con la calculada a partir de las coordenadas geográficas de las empresas. Abordamos el problema incorporando un conjunto de datos geojson de los Estados Unidos, centrado en los Estados criterio
(“New Jersey”, “California”, “Florida”, “Illinois”).
Implementamos la creación de un GeoData Frame usando las coordenadas de las empresas y realizamos una unión espacial con el conjunto de datos de los Estados criterio. Esta estrategia corrigió la asignación incorrecta de estados, mejorando significativamente la precisión de la información geográfica. La corrección asegura una representación más fiel de la ubicación de las empresas en nuestro análisis.

3. Pipeline ETL automatizado: Este hito se refiere a la implementación de un proceso ETL que se ejecuta automáticamente. Esto puede implicar el uso de herramientas y tecnologías que permiten programar y automatizar el proceso ETL.

4. Diseño del Modelo ER: Este hito implica el diseño del modelo Entidad-Relación (ER) para la base de datos. El modelo ER es una representación gráfica de los datos en una base de datos, que muestra las entidades (tablas), los atributos (columnas) y las relaciones entre las entidades [5].

5. Pipelines para alimentar el DW: Este hito se refiere a la implementación de pipelines que se encargan de alimentar los datos al Data Warehouse. Estos pipelines pueden estar diseñados para cargar datos en intervalos regulares, en respuesta a eventos específicos, entre otros [4].

6. Data Warehouse: Este hito se refiere a la implementación de un Data Warehouse. Un Data Warehouse es una base de datos que almacena datos históricos y actuales de una organización, que se utilizan para el análisis y la toma de decisiones [5].

7. Automatización: Este hito se refiere a la automatización de tareas en el proceso de ETL. Esto puede implicar la automatización de la extracción de datos, la transformación de datos, la carga de datos, entre otras tareas [1].

8. Validación de datos: Este hito se refiere a la validación de los datos para asegurar su calidad. Esto puede implicar la verificación de la integridad de los datos, la comprobación de la consistencia de los datos, la detección y corrección de errores en los datos, entre otras tareas [1].

9. Documentación: Este hito se refiere a la documentación de todo el proceso de ETL. Esto puede implicar la documentación de los requisitos del proyecto, la documentación de las decisiones tomadas durante el desarrollo del proyecto, la documentación de las herramientas y tecnologías utilizadas, entre otras cosas [4].

10. Diagrama ER detallado (tablas, PK, FK y tipo de dato): Este hito se refiere a la creación de un diagrama ER detallado. Este diagrama debe incluir todas las tablas en la base de datos, las claves primarias (PK) y las claves foráneas (FK) en cada tabla, y el tipo de dato de cada columna en la tabla [5].

11. Diccionario de datos: Este hito se refiere a la creación de un diccionario de datos. Este diccionario debe proporcionar información detallada sobre cada tabla, columna y tipo de dato en la base de datos [4].

Diccionario de datos

12. Workflow detallando tecnologías: Este hito se refiere a la creación de un flujo de trabajo que detalla las tecnologías utilizadas en el proyecto. Este flujo de trabajo debe incluir información sobre las herramientas y tecnologías utilizadas en cada etapa del proceso ETL .