# Amazon Recommender System Project

Integrantes:
- Juan Sebastián Sara
- Jose Chachi
- Bladimir

# Caso de estudio

- El análisis de reviews de usuarios en plataformas e-commerce, permiten a las empresas comprender mejor la percepción de sus productos y servicios
- Este estudio se centra en clasificar las reviews de Gift Cards de Amazon en el año 2023
- Se clasificaran como positivas y negativas
- Se utilizaran técnicas NLP y Machine Learning
- El objetivo es crear un modelo que sea capaz de identificar patrones y tendencias en los comentarios de los usuarios.

# Justificación

- Las opiniones de los usuarios son determinantes para la percepción de productos y servicios
- Permite a las empresas identificar patrones y tendencias
- Proporciona una forma estructurada y accesible de evaluar la calidad de los productos
- Alto grado de aprovechamiento del dataset para entrenar modelos de Machine Learning, garantizando una solución escalable y automatizada

https://www.google.com/url?sa=i&url=https%3A%2F%2Fespecializate.usat.edu.pe%2Fblog%2Fla-importancia-de-los-negocios-digitales%2F&psig=AOvVaw2Gt7GpL6_kUAQixKsSNLhe&ust=1733263638550000&source=images&cd=vfe&opi=89978449&ved=0CBQQjRxqFwoTCMD23a6MiooDFQAAAAAdAAAAABAE

# Dataset

El dataset Amazon Reviews 2023 es un recurso de gran escala que contiene reseñas de productos de Amazon junto con información extensa en metadatos y enlaces entre usuarios y productos. Este fue diseñado para la investigación en sistemas de recomendación e interpretación del lenguaje natural.

| Año | #Review | #User | #Item | #R_Token | #M_Token | #Domain | Timespan |
|---|---|---|---|---|---|---|---|
| 2013 | 34.69M | 6.64M | 2.44M | 5.91B | – | 28 | Jun'96 - Mar'13 |
| 2014 | 82.83M | 21.13M | 9.86M | 9.16B | 4.14B | 24 | May'96 - Jul'14 |
| 2018 | 233.10M | 43.53M | 15.17M | 15.73B | 7.99B | 29 | May'96 - Oct'18 |
| 2023 | 571.54M | 54.51M | 48.19M | 30.14B | 30.78B | 33 | May'96 - Sep'23 |

# Tamaño del Dataset

Dataset de Gift Cards (Reviews)

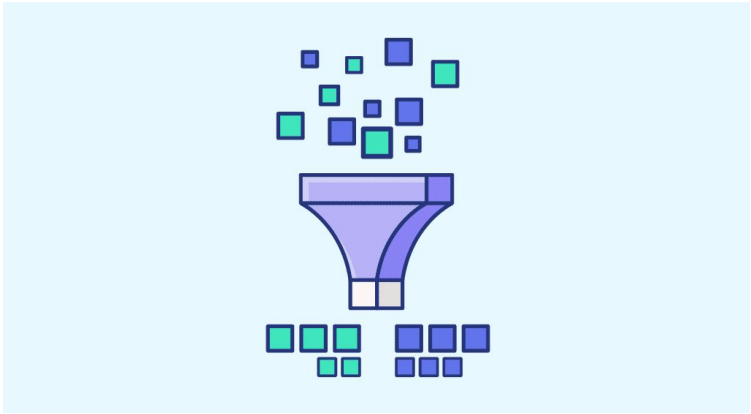
- Tamaño aprox. por línea: 233 B
- Cantidad de líneas: 152,410
- Peso total estimado: ~35.5 MB

Dataset de Gift Cards (Meta data)

- Tamaño aprox. por línea: 2.33 KB
- Cantidad de líneas: 1137
- Peso total estimado: ~2.65 MB

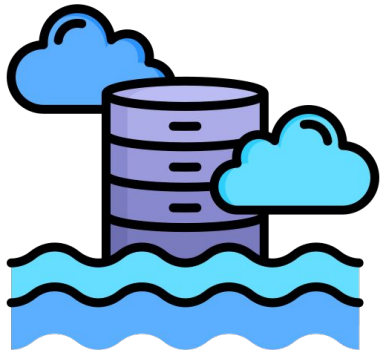
# Dificultades técnicas

- Limpieza de datos
- Desbalanceo del dataset
- Manejo de lenguaje o distintos idiomas
- Implementación del modelo LLM de Huggin Face
- El modelo LLM permite solo 512 tokens de texto



# Herramientas

mongoDB®

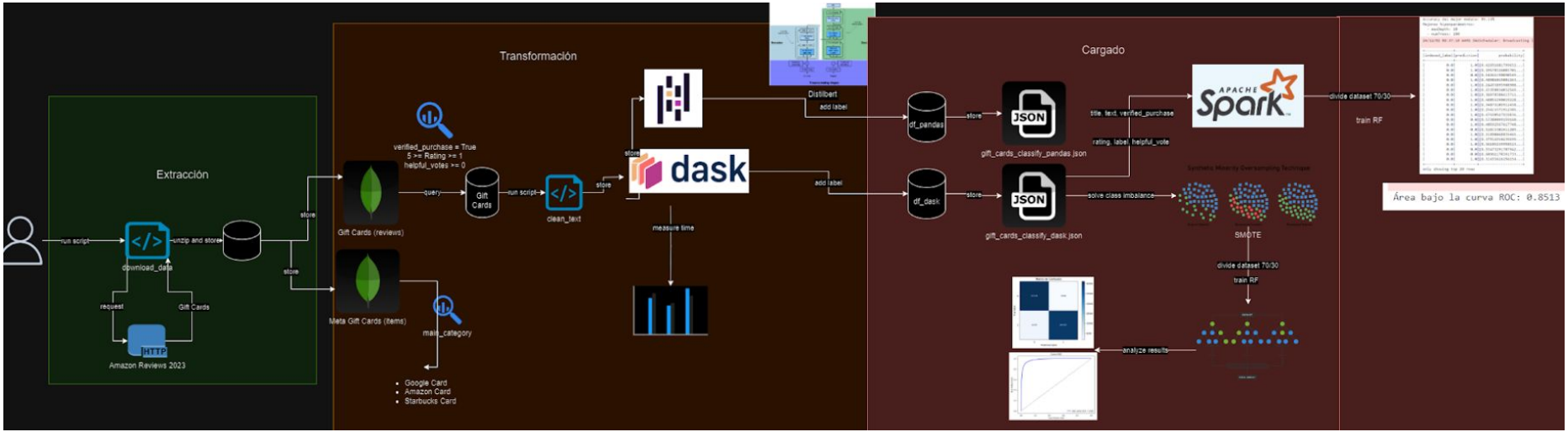


W&B

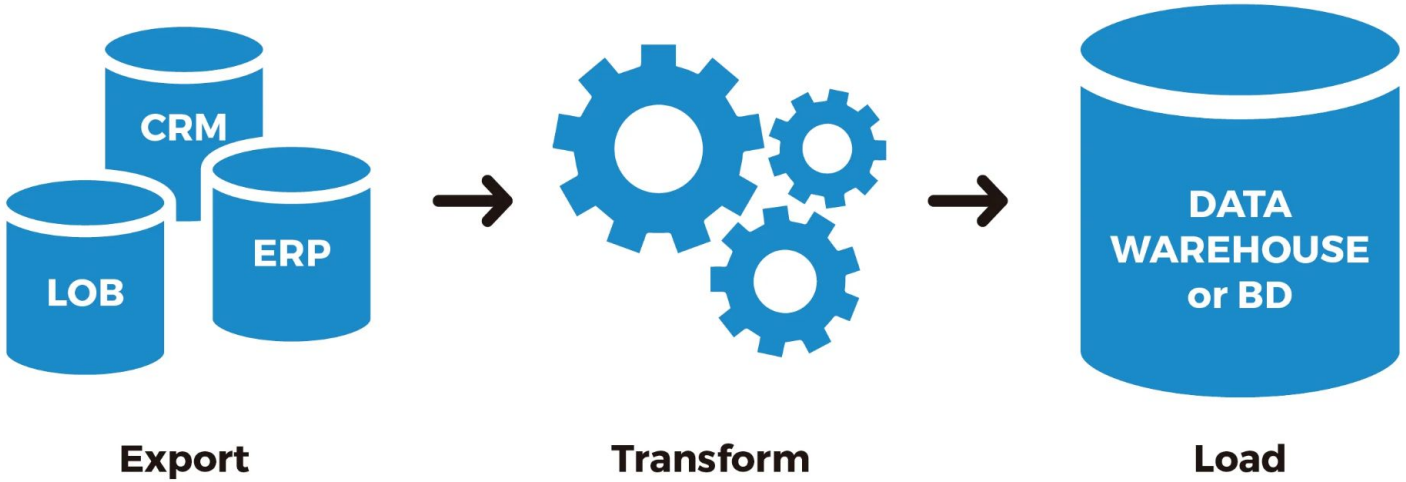
dask


APACHE Spark™

# Arquitectura del proyecto

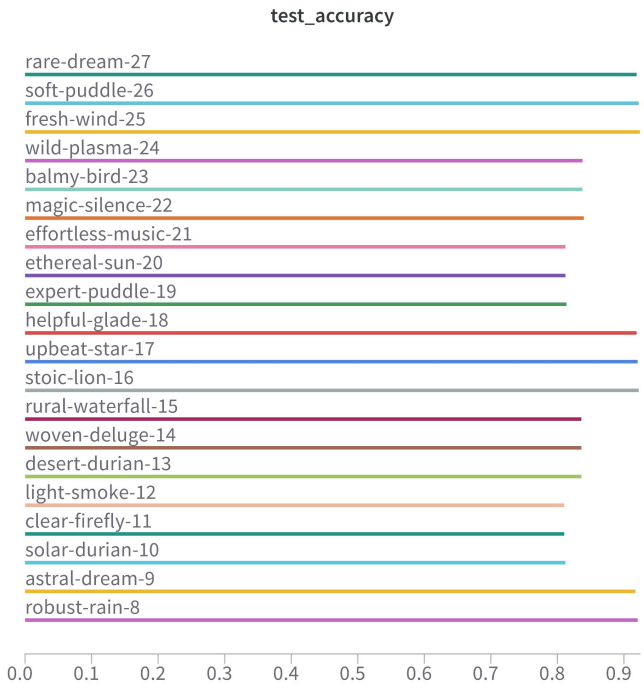


# ETL



# Resultados obtenidos
# Test accuracy

De todos los experimentos, el fresh-wind-25 tuvo el de mejor desempeño con un accuracy de 92.38%, y usando como base los siguientes hiperparámetros, n_estimators: 200, max_depth: null, min_samples_split: 2



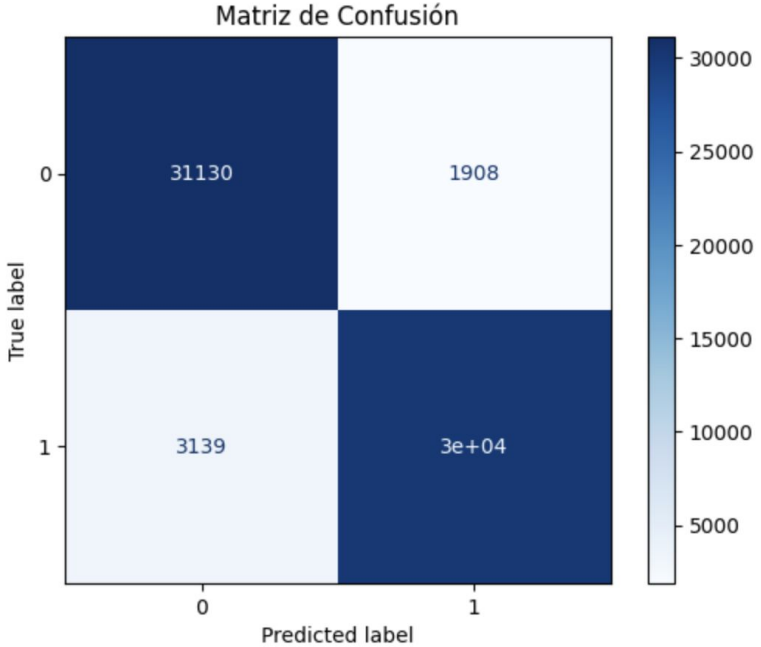
# Reporte de Clasificación

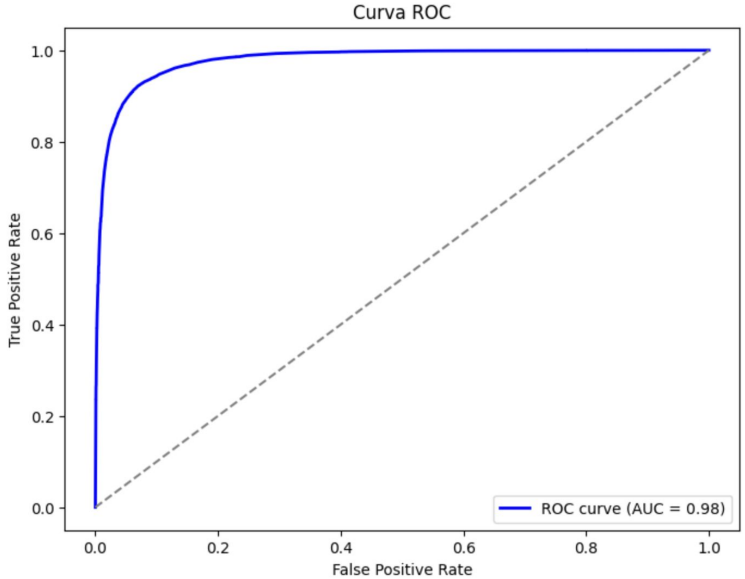
El modelo alcanzó un accuracy de 92.38% en los datos de prueba. La clase 0 (reviews negativas) obtuvo una precisión del 91% y un recall del 94%, mientras que la clase 1 (reviews positivas) alcanzó una precisión del 94% y un recall del 91%. Indicando un buen desempeño de predicción para ambas clases.

Accuracy: 0.9238
Reporte de clasificación:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.91 | 0.94 | 0.93 | 33038 |
| 1 | 0.94 | 0.91 | 0.92 | 33205 |
| accuracy | | | 0.92 | 66243 |
| macro avg | 0.92 | 0.92 | 0.92 | 66243 |
| weighted avg | 0.92 | 0.92 | 0.92 | 66243 |

# Matriz de confusión y curva ROC



# Posibles mejoras

- Ajustar los hiperparámetros con validación cruzada para evitar sobreajuste y mejorar la generalización del modelo.
- Mejorar el balanceo de clases, aplicando smote solo a características numéricas y para la parte de texto usar técnicas como resampling basado en tokens (Word2Vec, GPT).
- Implementar un pipeline para limpieza de texto que incluya detección de idiomas, para mejorar el manejo de textos en varios idiomas.

# Conclusiones

- Se logró crear un modelo que sea capaz de clasificar las reviews de los usuarios de manera eficiente y con un alto nivel de precisión
- Se cumplieron con los requisitos de ETL (extraccion, transformacion y carga) los datos, asegurando que estos se encuentren aptos para el entrenamiento del modelo
- La herramienta Dask es más rápido que Pandas al momento de limpiar la data, sin embargo esto no se vio al momento de correr el modelo LLM, ambos teniendo tiempos casi similares