



Proyecto: Análisis de costo

Para estimar los costos de desplegar el análisis de clustering con **KMeans** y la generación de **wordclouds** en producción, tomaremos en cuenta los siguientes factores:

1. Almacenamiento en la nube

- **Dataset de posts (2 GB) y Dataset de comentarios (13 GB).**
- En total, el almacenamiento es de **15 GB**.

Herramienta en la nube: Amazon S3

- **Costo estimado:** \$0.023 por GB al mes.

Referencia:

Precios Amazon Web Service S3 | Amazon Simple Storage Service
Descubra los precios Amazon Web Services S3 y pague solo por lo que utilice sin cuota mínima. Realice una estimación de su factura mensual con la calculadora AWS Amazon. Explore aquí para conocer el sistema de



<https://aws.amazon.com/es/s3/pricing/>



- **Costo mensual para 15 GB:** $15 \text{ GB} * \$0.023 = \$0.345/\text{mes}$.

2. Procesamiento de datos con Databricks

Para el análisis con **KMeans** (creación de 3 clusters) y la generación de **wordclouds**, se puede utilizar una plataforma de Big Data como Databricks:

Herramienta en la nube: Databricks (en AWS).

Estimación del tiempo de procesamiento: se supondrá que el procesamiento de datos y la generación de clusters toma aproximadamente 1 hora.

- **Tipo de instancia:** **Databricks Classic Jobs / Classic Jobs Photon clusters** (nodo de trabajo de bajo costo para procesamiento no interactivo).
- **Costo por hora:** Aproximadamente **\$0.15** por nodo de trabajo.
- **Nodos necesarios:** Para procesar 15GB de datos, es razonable utilizar **2 nodos**.
- **Costo de procesamiento:** $2 \text{ nodos} * \$0.15/\text{hora} * 1 \text{ hora} = \$0.30 \text{ por procesamiento}$.

Referencia:



3. Machine Learning con MLlib (KMeans)

- **KMeans** es una técnica computacionalmente intensiva, y su costo depende del uso de **Spark MLlib** en Databricks para realizar el clustering de los datos.
- **Costo adicional** para el uso de **Databricks ML**: no hay sobre costo adicional.

4. Generación de WordClouds

- Una vez generados los 3 clusters, se procederá a generar los **wordclouds**.
- El procesamiento para la generación de los wordclouds es menos intensivo en comparación con KMeans, por lo que se puede suponer que este proceso se completará en **2 minutos**.
- Utilizando los mismos nodos de procesamiento (2 nodos).
- **Costo para la generación de WordClouds**: $2 \text{ nodos} * \$0.15/\text{hora} * 0.03 \text{ horas} = \0.009 .