Tecnológico de Monterrey

Análisis de grandes volúmenes de datos

TC4034.10

Proyecto | lectura, escritura, archivos de Big Data PySpark

Equipo 37:

Alejandro Díaz Villagómez	A01276769
Alonso Pedrero Martínez	A01769076
César Iván Pedrero Martínez	A01366501
Emiliano Saucedo Arriola	A01659258

Docentes:
Dr. Iván Olmos Pineda
Mtra. Verónica Sandra Guzmán de Valle
Mtro. Alberto Daniel Salinas Montemayor

Fecha de entrega:
27 de abril del 2025

Descripción del tema de interés

El presente proyecto tiene como objetivo principal analizar las reseñas de productos publicadas en Amazon utilizando herramientas de Big Data, específicamente PySpark. Amazon, como empresa de reconocimiento mundial tanto en el ámbito del comercio electrónico como en el desarrollo de soluciones tecnológicas, destacando su plataforma AWS (Amazon Web Services), constituye un referente global en el sector. Además, su posicionamiento como líder indiscutible del mercado de marketplaces la convierte en una fuente valiosa de datos masivos para análisis de comportamiento del consumidor.

La selección de este tema responde a intereses tanto profesionales como personales. Como programadores y consultores de software, frecuentemente trabajamos con clientes que solicitan el desarrollo de plataformas de venta en línea, mostrando un creciente interés en conocer la percepción y aceptación de sus productos en el mercado digital. En este contexto, resulta fundamental contar con herramientas que permitan identificar tendencias, analizar sentimientos y extraer patrones ocultos en las opiniones de los usuarios.

Por tanto, el objetivo general de este proyecto es construir una solución que, a partir del procesamiento y análisis de grandes volúmenes de reseñas, permita:

- Identificar el sentimiento general asociado a los productos (positivo, negativo o neutral).
- Detectar tendencias y patrones recurrentes en las opiniones de los consumidores.
- Apoyar en la toma de decisiones estratégicas de marketing y mejora de productos, proporcionando información valiosa y procesable a partir de los datos analizados.

Este enfoque no solo fortalece nuestras competencias técnicas en Big Data (e incluso en temas como procesamiento de lenguaje natural - NLP), sino que también aporta una herramienta práctica y directamente aplicable a nuestro entorno laboral actual.

Selección del dataset

Datos generales

Nombre del dataset	Amazon-reviews
Enlace	https://www.kaggle.com/datasets/machharavikiran/amazon-reviews?resource=download
Publicador	Machha Ravi Kiran
Última actualización	Hace aproximadamente 2 años
Formato - Tamaño	CSV - 3.68 GB
Fuente	Repositorio Kaggle (acceso público)

El dataset seleccionado cumple con los criterios establecidos para este proyecto, ya que está directamente relacionado con el análisis de reseñas de productos, presenta un tamaño

adecuado para técnicas de Big Data, y su formato tabular favorece su procesamiento mediante PySpark. Además, la procedencia del dataset, proveniente de una fuente reconocida como Kaggle, garantiza además su calidad y fiabilidad para los fines académicos y profesionales de este trabajo.

Análisis exploratorio con PySpark

Para el análisis de este proyecto, se seleccionó el conjunto de datos "Amazon Reviews" disponible en la plataforma Kaggle. Utilizando PySpark, se procedió a la apertura y exploración preliminar de la base de datos.

La estructura del conjunto de datos es la siguiente:

- Número de registros: 6,906,564
- Número de columnas: 16

La descripción detallada de las columnas es la siguiente:

Columna	Tipo de dato	Descripción
marketplace	string	Mercado donde se realizó la compra
customer_id	integer	Identificador único del cliente
review_id	string	Identificador único de la reseña
product_id	string	Identificador del producto
product_parent	integer	Identificador del producto principal
product_title	string	Nombre del producto
product_category	string	Categoría del producto
star_rating	integer	Calificación otorgada por el cliente
helpful_votes	integer	Número de votos útiles recibidos por la reseña
total_votes	integer	Número total de votos recibidos
vine	string	Participación en el programa Vine de reseñas (Y para sí, N para no)
verified_purchase	string	Indicador de compra verificada (Y para sí, N para no)
review_headline	string	Encabezado de la reseña
review_body	string	Cuerpo del texto de la reseña
review_date	date	Fecha en que se realizó la reseña
sentiment	integer	Valor de sentimiento asignado a la reseña (1 para positivo, 0 para negativo)

Tras realizar una revisión detallada del conjunto de datos seleccionado, se confirma que el dataset contiene información altamente relevante para los objetivos planteados en el proyecto. En particular:

- ***Contenido de reseñas***: El dataset incluye tanto el encabezado (*review_headline*) como el cuerpo completo de las reseñas (*review_body*), lo cual proporciona el material textual necesario para realizar análisis de sentimientos y procesamiento de lenguaje natural (NLP).
- ***Valoraciones explícitas***: La columna *star_rating* ofrece una valoración numérica (de 1 a 5 estrellas) que puede ser utilizada como etiqueta supervisada para clasificar el sentimiento (positivo, neutro o negativo) en tareas de modelado predictivo.
- ***Datos adicionales para análisis contextual***: Variables como *product_category*, *verified_purchase* y *helpful_votes* permiten enriquecer el análisis, facilitando la segmentación de los datos por categoría de producto o validando la autenticidad de las opiniones emitidas.
- ***Volumen adecuado***: Con más de 6.9 millones de registros y 16 columnas, el volumen de datos satisface ampliamente los requisitos de trabajo con tecnologías de Big Data, permitiendo la implementación de técnicas de procesamiento distribuidas a gran escala mediante PySpark.
- ***Calidad de los datos***: El análisis preliminar indica que no existen valores faltantes en el conjunto de datos, lo que favorece la robustez de los modelos y análisis posteriores.
- ***Sesgo positivo en el sentimiento***: El análisis de la variable *sentiment* confirma una alta proporción de opiniones positivas, hecho que deberá ser considerado en las etapas de modelado para evitar sesgos en los resultados.
- ***Homogeneidad geográfica y de categorí***a: La totalidad de las reseñas procede del marketplace de Estados Unidos (marketplace = US) y pertenece a la categoría PC, lo que proporciona un enfoque específico pero también limita la generalización a otros mercados o categorías.
- ***Distribución uniforme de identificadores***: Variables como *review_id* y *product_id* presentan una distribución que garantiza unicidad y suficiente variabilidad, permitiendo análisis individuales o agregados de productos y reseñas.

El conjunto de datos seleccionado resulta plenamente adecuado para el desarrollo del proyecto, ya que proporciona tanto el contenido textual como las variables auxiliares necesarias para cumplir los objetivos de análisis de sentimiento y detección de patrones de comportamiento del consumidor en plataformas de comercio electrónico.

Asimismo, es importante destacar que, de acuerdo con la información disponible, ni los autores del conjunto de datos ni la comunidad de usuarios de Kaggle han reportado problemas relacionados con la calidad, integridad o estructura del mismo, lo que refuerza su idoneidad para los fines de esta investigación.

El código fuente detallado utilizado en este proyecto se encuentra disponible en el apartado de Anexos.

Anexos

Anexo 1. Código fuente

El código completo y detallado utilizado para el procesamiento y análisis del conjunto de datos puede consultarse en el siguiente enlace: *archivos de Big Data PySpark_Equipo37.ipynb*