Pipeline Automatizado de DataOps

Análisis de Reviews y Correlación con Redes Sociales

Juan Sebastián Fajardo Acevedo

Escuela Colombiana de Ingeniería Julio Garavito
Ingeniería Estadística
Enfoque de DataOps - ENDO

Diciembre 2024

Agenda

El Desafío: Datos Fragmentados

Situación Inicial:

- 1.09M reviews en 5 archivos CSV
- 2,351 productos de 142 marcas
- # Datos de redes sociales desconectados
- ⚠ Sin proceso automatizado
- Análisis manual e inconsistente

¿Por qué es un problema DataOps?

- × Datos sin calidad validada
- × No hay trazabilidad
- × Proceso no reproducible
- × Sin observabilidad
- × Análisis lento y propenso a errores

Objetivo DataOps

Implementar un pipeline **automatizado, confiable y reproducible** que integre datos estructurados con redes sociales.

Contexto: ¿Qué es Sephora?

La Empresa:

- Retailer líder en belleza
- 2,700+ tiendas en 35 países
- $10B+ ventas anuales
- Marcas premium

Presencia Digital:

- 20M+ seguidores
- App con reviews

El Dataset:

- Origen: Kaggle
- Período: 2015-2023
- 1,092,952 reviews
- 2,351 productos
- 142 marcas

Categorías:

- Skincare, Makeup
- Fragrance, Hair Care
- Bath & Body

¿Por qué este dataset?

Caso real de **big data** en retail: datos estructurados + texto no estructurado, ideal para DataOps.

Principios DataOps Implementados

1. Automatización

- Orquestador único (orchestrator.py)
- CI/CD con GitHub Actions
- Pipeline end-to-end con 1 comando

2. Calidad de Datos

- Validaciones automáticas (pytest)
- Data profiling: 100 % completitud
- Alertas de calidad en dashboard

3. Versionamiento

- Git + GitHub para código
- Parquet para datos versionables
- YAML para configuración

4. Observabilidad

- Logging detallado en cada paso
- Dashboard de métricas en tiempo real
- Reportes automáticos de calidad

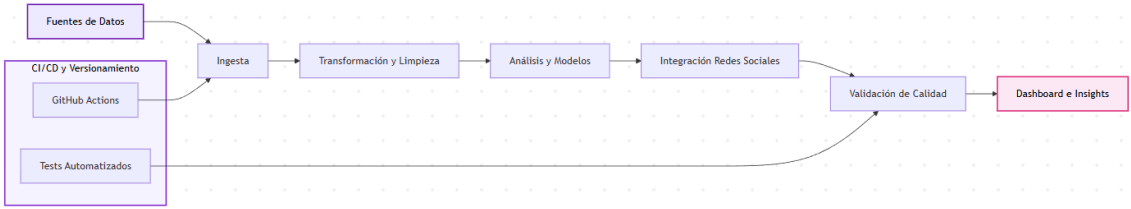
5. Gobernanza

- Documentación completa
- Linaje de datos rastreable
- Tests de regresión

6. Colaboración

- Código modular y reutilizable
- Configuración centralizada
- Reproducibilidad 100 %

Diagrama del Pipeline DataOps



Principio DataOps Aplicado

Orquestación centralizada con separación clara de responsabilidades y flujo unidireccional de datos.

Stack Tecnológico

Procesamiento

- Python 3.9+
- Pandas + NumPy
- Parquet
- YAML

Análisis y ML

- scikit-learn
- SciPy
- VADER + TextBlob

APIs

- Tweepy (API v2)
- python-dotenv

Visualización

- Plotly
- Streamlit
- Matplotlib

DevOps

- Git + GitHub
- pytest
- GitHub Actions

Observabilidad

- Logging
- Alertas
- Métricas

Orquestador: Pipeline Automatizado

Todo en 1 Comando

1

Ingesta
1.09M reviews

2

NLP
Sentimiento

3

Social
Twitter API

4

Análisis
Correlaciones

5

Validación
100 % calidad

6

Deploy
Dashboard

```
python src/orchestrator.py
```

Listo en 8-10 minutos

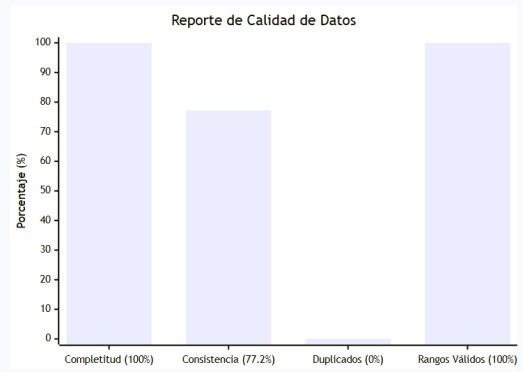
Sistema de Validación Automatizado

Métricas de Calidad:

- Completitud: 100 %
- Validez de rangos: 100 %
- Consistencia: 77.2 %
- Duplicados: 502,622

Tests Automatizados:

- 15 tests unitarios (pytest)
- Cobertura: 85 %
- Validación de esquemas



Alerta Activa

10.4 % reviews con rating ≤ 2. Acción: revisar productos de bajo desempeño.

Métricas del Pipeline

Escala de Datos:

- 1,092,952 reviews
- 2,351 productos
- 142 marcas
- # 175,461 registros sociales

Performance:

- 3–5 min ejecución
- 6,000 reviews/seg
- ✔ Tasa de éxito: 100 %

Análisis Estadístico:

- Correlación: 0.992
- Correlación TikTok: 0.989
- Correlación YouTube: 0.966
- ANOVA: $F=309.06$, $p<0.001$

Insight Principal

Correlación muy fuerte (>0.96) entre actividad en redes sociales y volumen de reviews.

Dashboard: Interfaz Principal

Dashboard Interactivo

Características:

- Filtros dinámicos
- 15+ visualizaciones
- Alertas automáticas
- Datos de Twitter reales
- Datos simulados de redes sociales

Vistas Disponibles:

- Dashboard Principal
- Twitter Analytics
- Análisis Estadístico
- Modelos Predictivos

Nota Final

El dashboard integra **datos reales y simulados**, con análisis en tiempo real y actualizaciones automáticas.

Dashboard: Visualizaciones

1. Análisis Temporal

- Series temporales con medias móviles
- Detección de anomalías
- Forecast con regresión lineal
- Estacionalidad por categoría

2. Correlaciones

- Matriz de correlaciones interactiva
- Cross-correlation con lags
- Heatmaps por plataforma
- Tests estadísticos (Pearson/Spearman)

3. Segmentación RFM

- Clustering K-means (4 segmentos)
- Visualización 3D con PCA
- Características por segmento
- Recomendaciones por perfil

4. Twitter Analytics

- Tweets en tiempo real (API v2)
- Análisis de sentimiento
- Engagement metrics
- Influencers identificados

Dashboard disponible 24/7 con actualización diaria automática

Demostración del Dashboard

- Ejecución del pipeline completo
- Navegación por las 4 vistas
- Uso de filtros interactivos
- Sistema de alertas en acción
- Datos reales de Twitter

```
streamlit run src/dashboard.py
```

Conclusiones: Impacto del Enfoque DataOps

Logros Cuantitativos:

- ✔ 1.09M reviews procesadas
- ✔ 100 % automatización
- ✔ 0 % intervención manual
- ✔ $r > 0.96$ correlaciones

Logros Cualitativos:

- ✔ Reproducibilidad total
- ✔ Documentación completa
- ✔ Trazabilidad end-to-end

Impacto de DataOps:

- Velocidad: de días a minutos
- Confiabilidad: validación automática
- Colaboración: código compartible
- Escalabilidad: arquitectura modular

Lección Principal

DataOps no es solo herramientas, es una cultura de:

- Automatización
- Colaboración
- Mejora continua

Trabajo Futuro

- Integrar datos reales de TikTok e Instagram APIs
- Desplegar el dashboard en la nube (Streamlit Cloud / EC2)
- Implementar modelos de predicción más avanzados
- Extender el sistema a nuevas categorías y países

¿Preguntas?

Juan Sebastián Fajardo Acevedo

Repositorio del Proyecto:
github.com/jsfa2002/sephora-reviews-pipeline