

Actividad Análisis Crítico de la Calidad y Optimización de Datos en Big Data

Objetivo General

Aplicar técnicas de limpieza de datos y evaluar el impacto de la transformación de un dataset en formato CSV a Parquet, destacando las mejoras obtenidas en términos de calidad, eficiencia y optimización en un entorno Big Data.

Descripción de la Actividad

A partir del trabajo realizado en el cuaderno de Google Colab, deberás ejecutar todas las fases del proceso y elaborar un informe técnico.

Enlace al cuaderno de Colab:

<https://colab.research.google.com/drive/19uqfqmOEj-pT-mk4EiZxv-fQlivWJIWI?usp=sharing> → importante una vez dentro guardar como copia en Drive.

Contenido del Informe (en PDF)

1. Análisis de la Limpieza de Datos

- Explica los principales errores detectados en el dataset original.
- Describe las técnicas aplicadas para corregirlos.
- Presenta las métricas pre y post limpieza.
- Reflexiona sobre el impacto de la mala calidad de los datos.

2. Análisis de la Transformación a Parquet

- Compara tamaño de archivo, tiempo de carga y eficiencia.
- Explica las ventajas del formato Parquet.
- Contextualiza el uso de CSV vs Parquet.

3. Conclusión Global

- Resume los beneficios de la limpieza de datos y la optimización con Parquet.
- Relaciona ambos procesos con escenarios reales de Big Data.

4. Evidencias

- Incluye capturas de pantalla clave de cada fase del proceso.

Entrega

Formato: Informe en PDF, nombrado como: Nombre_Apellido_Limpieza_Parquet.pdf

Criterios de Evaluación

Criterio	Peso
Análisis crítico de la limpieza de datos	40%
Evaluación de la transformación a Parquet	30%
Calidad del informe (claridad, estructura)	20%
Evidencias correctamente presentadas	10%

Notas Importantes

- No te limites a describir el proceso, analiza el impacto real.
- Relaciona los resultados con necesidades de eficiencia en proyectos reales.
- Reflexiona sobre la influencia en la toma de decisiones empresariales.