Informe de Proyecto

Subtítulo opcional

Eveliz Espinaco Milián
Richard Alejandro Matos Arderí

Universidad de La Habana,
Facultad de Matemática y Computación

20 de octubre de 2025

Resumen: Aquí va un breve resumen del informe.

Índice

1 Dataset Seleccionado

En esta sección se describe el conjunto de datos elegido para el desarrollo del proyecto.

1.1 Nombre, Fuente y Formato

- **Nombre:** Cambio climático: datos de temperatura de la superficie terrestre
- **Fuente:** https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data
- **Formato:** CSV

1.2 Variables Estadísticas

A continuación, se describen las variables incluidas en el dataset, detallando su significado, los valores que pueden tomar, y su clasificación correspondiente.

- **Fecha:** Indica el año de la medición. Los valores comienzan en 1750 para la temperatura media de la tierra y en 1850 para las temperaturas máximas y mínimas, así como para las temperaturas globales de océanos y tierra.
  Valores: Años (1750 en adelante).
  Tipo: Variable cuantitativa, escala ordinal.
- **País:** País donde se realizó la medición.
  Valores: Nombres de países (por ejemplo, España, México, Argentina).
  Tipo: Variable cualitativa, escala nominal.
- **Ciudad:** Ciudad específica de la estación meteorológica.
  Valores: Nombres de ciudades (por ejemplo, Madrid, Buenos Aires).
  Tipo: Variable cualitativa, escala nominal.
- **Latitud:** Coordenada geográfica norte-sur de la estación.
  Valores: Números reales en grados decimales (por ejemplo, 40.4168).
  Tipo: Variable cuantitativa, escala de razón.
- **Longitud:** Coordenada geográfica este-oeste de la estación.
  Valores: Números reales en grados decimales (por ejemplo, -3.7038).
  Tipo: Variable cuantitativa, escala de razón.
- **Temperatura media del terreno:** Representa la temperatura media global de la superficie terrestre, expresada en grados Celsius.
  Valores: Números reales (por ejemplo, 13.5°C).
  Tipo: Variable cuantitativa, escala de razón.
- **Incertidumbre de la temperatura media del terreno:** Intervalo de confianza del 95 % alrededor del promedio de la temperatura media del terreno.
  Valores: Números reales positivos (por ejemplo, 0.12°C).
  Tipo: Variable cuantitativa, escala de razón.

2 Justificación del Dataset

El dataset seleccionado contiene registros históricos de temperatura de la superficie terrestre, recolectados a lo largo de varias décadas mediante distintos métodos e instrumentos. Su uso está justificado por la complejidad inherente a los datos, que exige una gran cantidad de limpieza, normalización y procesamiento para poder extraer conclusiones válidas sobre las tendencias climáticas a largo plazo. Los primeros registros fueron obtenidos por técnicos que utilizaban termómetros de mercurio, donde incluso pequeñas variaciones en la hora de la medición podían alterar significativamente los valores registrados. Posteriormente, en la década de 1940, la construcción de aeropuertos obligó al traslado físico de muchas estaciones meteorológicas, introduciendo discontinuidades espaciales en las series de datos. Más adelante, en los años 80, se incorporaron termómetros electrónicos, los cuales presentan un sesgo sistemático de enfriamiento que debe ser corregido para garantizar la coherencia del análisis.

Este contexto histórico y técnico convierte al dataset en un excelente candidato para evaluar los conocimientos adquiridos en la asignatura de Procesamiento de Grandes Volúmenes de Datos. A pesar de que el volumen original es relativamente pequeño, la riqueza estructural, la heterogeneidad temporal y la presencia de sesgos lo hacen ideal para simular escenarios reales de Big Data, donde la calidad y la gobernanza de los datos son tan importantes como la cantidad.

2.1 Volumen

El dataset cuenta con aproximadamente 8.6 millones de registros, lo que lo convierte en un ejemplo representativo de escenarios reales de Big Data. Este volumen masivo permite aplicar técnicas avanzadas de procesamiento distribuido, como particionamiento, replicación y procesamiento paralelo, utilizando herramientas como Hadoop o Spark. La gran cantidad de datos facilita la segmentación por décadas, estaciones meteorológicas, países, o tipo de sensor, permitiendo diseñar esquemas de particionamiento que reflejan los principios de escalabilidad horizontal. Además, el tamaño del dataset posibilita la realización de análisis estadísticos robustos, la detección de patrones complejos y la generación de modelos predictivos con mayor precisión. La integración con fuentes externas (como altitud, ubicación geográfica, eventos históricos) y la generación de datos derivados amplían aún más las posibilidades de exploración y simulación en entornos de procesamiento masivo.

2.2 Características

Los atributos presentes en el dataset incluyen fecha de medición, ubicación geográfica, tipo de instrumento utilizado, temperatura registrada, y metadatos asociados a la estación meteorológica. La temporalidad es extensa, abarcando desde principios del siglo XX hasta la actualidad, lo que permite estudiar fenómenos de largo plazo como el cambio climático, la variabilidad estacional, y los efectos de urbanización. El dataset no está etiquetado en el sentido clásico de aprendizaje supervisado, pero permite generar etiquetas derivadas (por ejemplo, anomalías térmicas, zonas de cambio abrupto, o eventos extremos) mediante procesamiento. El nivel de ruido es alto: hay inconsistencias en las

unidades, valores faltantes, duplicados, y sesgos sistemáticos que deben ser corregidos. Esta situación obliga a aplicar técnicas de limpieza, imputación, normalización y reconciliación de fuentes, lo cual es central en el estudio de grandes volúmenes de datos.

2.3 Pertinencia

El dataset se relaciona directamente con los objetivos de la asignatura, ya que permite aplicar de forma integrada todos los conceptos clave: ingestión de datos desde múltiples fuentes, limpieza intensiva, transformación distribuida, almacenamiento optimizado, y análisis escalable. Además, su temática —las tendencias climáticas— es de alta relevancia social y científica, lo que motiva el trabajo y permite conectar la teoría con problemas reales. El proyecto puede incluir la simulación de un Data Lake con zonas raw, curated y analytics; el uso de herramientas como Apache Spark para agregaciones por década; y la visualización de resultados mediante dashboards interactivos. En conjunto, el dataset ofrece una oportunidad única para evaluar competencias técnicas, analíticas y metodológicas en un entorno controlado pero representativo de los desafíos del procesamiento de grandes volúmenes de datos.