**Introducción**

Este proyecto final del primer curso de Big Data e Inteligencia Artificial se centra en el proceso completo de análisis de datos: desde su ingesta, almacenamiento y procesamiento, hasta su visualización e interpretación.

Para desarrollar este estudio, hemos elegido la climatología como temática, enfocándonos específicamente en las precipitaciones en Australia entre los años 2008 y 2017. (Los archivos han sido sacado del siguiente enlace: <https://www.kaggle.com/datasets/trisha2094/weatheraus>).

Nuestro objetivo es extraer y analizar métricas clave que permitan comprender mejor las tendencias climáticas en este periodo.

Las métricas principales que evaluaremos son:

* Fecha con la temperatura más alta registrada.
* Fecha con la temperatura más baja registrada.
* Fecha con la mayor cantidad de lluvia registrada.
* Fecha con el viento más extremo registrado.
* Periodo de sequía más largo registrado.
* Cantidad de días sin lluvia por año.
* Media anual de lluvia y temperatura.
* Variabilidad de temperatura máxima y mínima por año.
* Días que superaron los 40°C por año (temperatura extrema).
* Días con temperaturas menores a -5°C por año *(en Australia, el frío extremo es poco frecuente, por lo que establecemos este umbral como referencia)*.
* Días con viento extremo (>93 km/h) por año.

**Ingesta y almacenamiento de datos**

Para iniciar este este proyecto lo primero que tenemos que hacer es la ingesta de datos, para ello vamos a usar S3 para la entrada y el almacenamiento de archivos, al igual que usaremos en su trasformación y filtración de datos Glue, después lo volveremos a volcar con los datos que requerimos en el S3 y realizaremos las consultas con colab.

También trataremos los datos con un ciclo de vida para que no ocupe más recursos de los que se debe.

**Inserción de datos**:

Lo primero que se va a realizar en la inserción de datos, para ello vamos al S3 y crearemos un bucket de entrada de datos (nombre dado finalentrada):

Una captura de pantalla de una computadora

El contenido generado por IA puede ser incorrecto.

Una captura de pantalla de una computadora

El contenido generado por IA puede ser incorrecto.

Subimos con los archivos que vamos a trabajar:

Una captura de pantalla de una computadora

El contenido generado por IA puede ser incorrecto.

(se suben 3 archivos por los experimentos durante el trabajo)

Tras subir los archivos y antes de hacer nada, tocaremos el ciclo de vida. Para ello lo primero que tenemos que hacer es habilitar el control de versiones para evitar problemas.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

El contenido generado por IA puede ser incorrecto.

Dado que estos datos no son de alta prioridad, pero queremos conservarlos durante el mayor tiempo posible sin ocupar recursos innecesarios, configuraremos una política de ciclo de vida en que no se borren durante 10 años. Esto nos permitirá optimizar el almacenamiento, asegurando que los datos permanezcan accesibles, pero con un uso eficiente de los recursos disponibles. Interfaz de usuario gráfica, Texto, Correo electrónico

El contenido generado por IA puede ser incorrecto.Como es un archivo de gran duración, lo dejaremos durante un año presente, tras pasar dicho año, se desplazara a s3glacier flexible retrival. Tras pasar 10 años, se eliminará.

Una captura de pantalla de una computadora

El contenido generado por IA puede ser incorrecto.

Una captura de pantalla de una red social

El contenido generado por IA puede ser incorrecto.

Tras crear la reglade vida, ya pasamos a generar otro bucket de salida de datos que lo usaremos más adelante:

Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.

El nombre dado es finalsalida.

**Procesamiento de datos con AWS Glue**

Una vez configurado el almacenamiento y el ciclo de vida de los datos, es momento de proceder con su transformación y filtrado mediante AWS Glue. Este servicio nos permite automatizar la extracción y preparación de los datos almacenados en S3, optimizando su estructura para posteriores consultas y análisis.

El primer paso en este proceso fue la creación de un rastreador dentro de AWS Glue, el cual se encargó de localizar y catalogar los archivos previamente almacenados. Para ello, establecimos un rastreador con el nombre "climatología", asegurándonos de que estuviera correctamente vinculado al bucket de entrada. Con este rastreador configurado, AWS Glue identificó los archivos disponibles en S3,

Una captura de pantalla de una computadora

El contenido generado por IA puede ser incorrecto.

Le daremos el nombre de climatología.

Una captura de pantalla de una computadora

El contenido generado por IA puede ser incorrecto.

Elegimos el archivo S3 que vamos a trabajar (después se modificaría esto para detectar la carpeta)

Interfaz de usuario gráfica, Texto, Aplicación

El contenido generado por IA puede ser incorrecto.

Una captura de pantalla de una computadora

El contenido generado por IA puede ser incorrecto.

Una captura de pantalla de una computadora

El contenido generado por IA puede ser incorrecto.

Creamos y dejamos listo:

Una captura de pantalla de una computadora

El contenido generado por IA puede ser incorrecto.

Una vez completada la ejecución del rastreador, accedemos al visualizador de tablas de AWS Glue, donde podemos examinar los datos detectados antes de proceder con su transformación final. En esta etapa, se verificó la correcta delimitación de los datos, asegurándonos de que estuvieran estructurados de manera clara y consistente. Dado que el archivo contenía información que no consideramos relevante para el análisis, realizamos una limpieza y filtrado, eliminando columnas innecesarias y organizando los registros con los parámetros específicos que necesitamos para la extracción de métricas.

El procedimiento es: primero colocar una entrada de S3, eligiendo el archivo que vamos a utilizar.

Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.

Para poder separar los datos en tablas legibles, hemos tenido que poner la delimitación con la (,)

A continuación, como dicho archivo tiene muchos datos que no los consideramos necesarios, los eliminaremos de la tabla al igual que pondremos los datos correspondientes a dicho dato:

Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.

Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.

Interfaz de usuario gráfica, Texto

El contenido generado por IA puede ser incorrecto.

Con los datos ya preparados y ajustados, procedimos a volcarlos en el bucket de salida previamente configurado, "finalsalida".Una captura de pantalla de una red social

El contenido generado por IA puede ser incorrecto.

Guardamos y corremos.

Una captura de pantalla de una computadora

El contenido generado por IA puede ser incorrecto.

Una captura de pantalla de una computadora

El contenido generado por IA puede ser incorrecto.

Revisamos que tenemos el archivo listo en el S3.

Una captura de pantalla de una computadora

El contenido generado por IA puede ser incorrecto.

Este paso nos permitió contar con un conjunto de datos refinado, listo para ser consultado y analizado en profundidad. Gracias a AWS Glue, pudimos automatizar el proceso de identificación, limpieza y almacenamiento, optimizando el flujo de trabajo y asegurando que los datos sean accesibles en su formato más útil para la fase final de visualización en Google Colab.

**Visualización de datos**

Para importar los datos desde S3, nos encontramos con la limitación del rol de laboratorio, que impide la importación directa desde Colab. Como alternativa, optamos por descargar manualmente los archivos y trabajar con ellos localmente en Colab. Este método nos permitió acceder a la información sin restricciones y proceder con el análisis de datos.

Interfaz de usuario gráfica, Texto, Aplicación

El contenido generado por IA puede ser incorrecto.

Abrimos el colab.

Interfaz de usuario gráfica, Texto, Aplicación

El contenido generado por IA puede ser incorrecto.

Importamos en el S3 el archivo que hemos descargado.

Importamos y previsualizamos los archivos para ver que se han cargado correctamente (se tubo varios problemas de visualización y se tubo que indagar como poder verlos sin cortar ni alteraciones)

Interfaz de usuario gráfica

El contenido generado por IA puede ser incorrecto.

Una vez visto que los archivos funcionan, comenzaremos con nuestra exposición de datos:

Lo primero que hacemos es dar formato y medidas para evitar fallos a la hora de coger los datos. (también se tubo que investigar pues las formulas son bastante más complejas y son muchos datos a analizar)

Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.

Tras tener los esquemas preparados, vamos a sacar la información y las medias que nos interesa:

En el primer bloque sacaremos los días más extremos que hemos tenido tanto de calor, frio, lluvia y viento.

Texto

El contenido generado por IA puede ser incorrecto.

La fecha con la sequía más larga de la historia:

Texto

El contenido generado por IA puede ser incorrecto.

Cantidad de días sin lluvia por año.

Texto

El contenido generado por IA puede ser incorrecto.

Media de lluvia y temperatura por año, y variabilidad de temperatura máxima y mina por año.

Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.

Días que superaron los 40°C por año (temperatura extrema), temperaturas menores a -5°C por año, y viento extremo (>93 km/h) por año.

Interfaz de usuario gráfica, Texto

El contenido generado por IA puede ser incorrecto.

Texto

El contenido generado por IA puede ser incorrecto.

**Conclusiones**

El análisis de datos climáticos en Australia entre 2008 y 2017 nos ha permitido observar patrones significativos en la variabilidad de temperaturas, precipitaciones y eventos extremos. Hemos identificado los días más cálidos, fríos y lluviosos del período, así como el día con vientos más intensos. También hemos determinado la duración de la sequía más larga registrada, evidenciando la recurrencia de períodos de escasez de lluvia.

Las medias anuales de lluvia y temperatura nos proporcionaron una perspectiva general del comportamiento climático en distintos años, ayudándonos a detectar posibles cambios en las tendencias de precipitaciones. La variabilidad de temperatura también mostró fluctuaciones considerables en las diferencias entre temperaturas máximas y mínimas, lo que podría tener implicaciones sobre la percepción térmica y el impacto en los ecosistemas.

Asimismo, los eventos climáticos extremos resaltan la importancia de monitorear el número de días con temperaturas superiores a 40°C y vientos extremos, factores que pueden influir en la vida cotidiana, la agricultura y la infraestructura del país.

Este análisis ha demostrado la utilidad del procesamiento de datos para comprender mejor la climatología y su evolución, estableciendo un marco para futuras investigaciones y posibles aplicaciones en la predicción de eventos climáticos.