Pre-Proyecto: Configuración del sistema Hadoop y estudio sobre la calidad del aire y su impacto en la salud

Por Carlos Andrés González Bono

Los datos y sus características

Los datos escogidos para este estudio se tratan de una tabla con datos sobre la calidad del Aire de Nueva York. Se trata de información pública, proporcionada por el ayuntamiento de Nueva York sobre la emisión y presencia de determinados químicos en el aire y la incidencia de ciertas enfermedades y muertes presuntamente relacionadas con éstos; los datos han sido recogidos entre 2005 y 2022 y se presentan en un archivo .csv con la siguientes columnas:

- Unique ID: Indentificador único de cada registro.
- Indicator ID: Identificador de las características de la medida (Cada combinación de Name, Measure y Measure Info tiene su ID).
- Name: Nombre de la medida.
- Measure: Tipo de medida.
- Measure Info: Información de medida.
- Geo Type Name: Tipo de información de localización.
- Geo Join ID: ID de localización.
- Geo Place Name: Localización (Hay más de una por 'Geo Join ID', deduzco que zonas cercanas pueden compartirlo).
- Time Period: Periodo que abarca o en el que ocurre la medida.
- Start_Date: Fecha de la medida.
- Data Value: Valor de la medida.
- Message: Notas (en este caso, totalmente vacío).

Al tratarse de un archivo .csv, a priori uno podría pensar que son datos estructurados, pero al haber datos de distintos tipos compartiendo columnas, se trata de datos muy poco homogéneos y que no se pueden operar entre sí, por lo que se podrían considerar semi-estructurados. Cualquier análisis requerirá estructuración y limpieza de estos datos.

Variables principales

Aunque ésto no se trate de un artículo periodístico, uno de los objetivos es la investigación, y se pueden usar las 5Ws para describir la información que vamos a manejar. Para abstraer los datos, manejarlos con menos sesgos y de una forma más sencilla, emplearé los campos "ID" para ordenar y manejar datos. Concretamente:

- ¿Qué? (What?)
  - Indicator ID: Cada ID corresponde a una combinación de Name, Measure y Measure Info, enlazando toda la información de la medida en una sola variable. Previamente tendré que identificar los IDs que correspondan a emisiones y presencia de químicos, y diferenciarlos de los IDs que indiquen incidencias clínicas.
  - Data Value: El valor de la medida.
- ¿Dónde? (Where?)
  - Geo Join ID: Cada ID corresponde a un área. Unas áreas pueden englobar a otras, y aunque a priori eso no tiene porqué generar un conflicto en los datos, es algo que tendré que identificar y considerar, mayormente consultando Geo Type Name.
- ¿Cuándo? (When?)
  - Start_Date: Fecha de la medida
  - Time Period: Periodo que abarca (o en el que ocurre) la medida.

El quién (Who?) son los habitantes del dónde (Where?) y el porqué (Why?) es uno de los objetivos del estudio. No debería haber dos medidas diferentes de la misma característica el mismo día y en el mismo lugar, por lo que parece que el Unique ID no nos va a aportar mucho, por ello está, a priori, descartado.

Objetivos del estudio

Los objetivos de este estudio son 3:

- Configurar un clúster real de Hadoop: Para consolidar conocimientos de arquitectura de Big Data y la integración de Hadoop con Spark.
- ¿Porqué? (Why?) - Descubrir y cuantificar la relación entre la presencia de contaminantes y su efecto en la salud (si existe).
- Desarrollar una hipótesis y ponerla a prueba con datos previamente no analizados (opcionalmente, si fuera posible, con Machine Learning).

Fuentes

- Air Quality | Kaggle : https://www.kaggle.com/datasets/jasmeet0516/air-quality
- Air Quality - Catalog (The City of New York) : https://catalog.data.gov/dataset/air-quality