Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Departamento de Ciencias de la Computación

Análisis de las temperaturas mundiales

Integrantes: Ariel Barraza
Valentina González
Maximiliano Vargas
Profesor: Aidan Hogan
Auxiliar: Alberto Moya
Ayudantes: Adriana Concha
Daniel D. Pinto
Gabriel Norambuena
Fecha de entrega: 21 de Junio de 2019

Índice de Contenidos

Lista de Figuras

Lista de Tablas

1. Objetivos

El principal objetivo del proyecto es realizar un análisis a un dataset que contiene la temperatura de la tierra en distintos lugares, con el fin de mostrar cuanto ha cambiado el clima con el paso del tiempo. Se quiere analizar en detalle el promedio de temperaturas por país.

Otro objetivo es utilizar y aprovechar las tecnologías que se enseñaron en el curso CC-5212 para realizar análisis de una gran magnitud de datos.

2. Datos

2.1. Origen

Los datos fueron recolectados por NOAA's Global Historical Climatology Network (GHCN). Estos datos fueron producidos gracias a resúmenes de clima de estaciones terrestres a través del mundo que han sido sometidos a un conjunto común de revisiones de control de calidad.

Los datos se obtuvieron a través de más de 20 fuentes, en donde se encuentran incluidos algunos datos de cada año desde 1763 hasta la fecha.

El volumen de los datos usados en este proyecto es el orden de 36 millones de filas (3 millomes de filas por año, por 12 años seleccionados). El dataset original tiene alrededor de 300 millos de filas).

2.2. Países

Se tiene un csv el cual contiene la siglas del país y el nombre del país. Los datos de una misma fila se encuentran separados por comas. En la siguiente figura se muestran algunos datos del archivo.

```
AC,Antigua and Barbuda
AE,United Arab Emirates
AF,Afghanistan
AG,Algeria
AJ,Azerbaijan
AL,Albania
AM,Armenia
AO,Angola
AQ,American Samoa [United States]
AR,Argentina
AS,Australia
AU,Austria
AY,Antarctica
BA,Bahrain
BB,Barbados
BC,Botswana
BD,Bermuda [United Kingdom]
```

Figura 1: Archivo csv países

2.3. Datos meteorológicos

Se tiene un csv el cual contiene la temperatura que registro una estación. El csv contiene los siguientes campos: (codigo-estacion, maxima-temperatura-del-dia, minima-temperatura-del-dia, temperatura-tiempo-observacion, precipitación, Nevada,profundidad de la nieve, otros-elementos).

En la siguiente figura se muestran algunos datos del archivo.

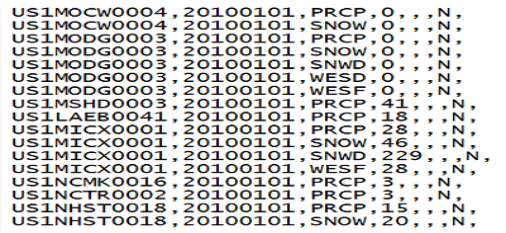

Figura 2: Archivo csv datos meteorológicos

2.4. Estaciones meteorológica

Se tiene un csv el cual contiene la información de la estación meteorológica en donde se hizo la medición. El csv contiene los siguientes campos: (codigo-estacion, latitud, longuitud, elevacion, nombre cuidad estacion, codigo-pais, pais).

En la siguiente figura se muestran algunos datos del archivo.

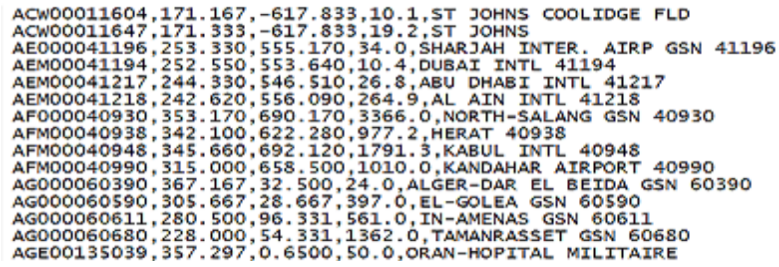

Figura 3: Archivo csv estaciones meteorológicas

3. Métodos y herramientas

Se realizará el análisis de los datos con la ayuda de Pig, la cual es una plataforma de alto nivel que permite crear programas MapReduce utilizados en Hadoop.

La metodología a seguir es realizar filter,join y group By para poder obtener la información que se necesita. Como se está trabajando con un gran volumen de datos, se debe filtrar primero las condiciones de los datos que se quieren analizar para luego realizar operaciones de map/reduce, lo que permite disminuir la cantidad de datos para realizar las otras operaciones.

3.1. Filtro de datos

Se tuvo que realizar una limpieza a los datos, ya que el objetivo era contrastar las temperaturas por los países y los CSV originales sólo tenían directamente acceso a los códigos de las estaciones meteorológicas.

Se realizó una filtración para extraer sólo parámetros de temperatura (y no precipitación por ejemplo), y luego una unión de tablas entre las mediciones y las estaciones con sus respectivos países. El código puede revisarse en el repositorio. Un ejemplo de filtración se muestra en la Tabla 1, donde en contraste con la Figura 2, sólo se muestran datos relacionados a temperaturas:

Tabla 1: Extracto de mediciones filtradas

USC00027281,20000101,TMAX,133	20000101	TMAX	133
USC00027281,20000101,TMIN,22	20000101	TMIN	22
USC00034988,20000101,TMAX,167	20000101	TMAX	167
USC00034988,20000101,TMIN,-17	20000101	TMIN	-17
USC00099466,20000101,TMAX,167	20000101	TMAX	167
USC00099466,20000101,TMIN,78	20000101	TMIN	78
USC00109601,20000101,TMAX,-22	20000101	TMAX	-22

3.2. Scripts

Se realizaron diversos scripts en PIG para realizar consultas interesantes:

- Promedios de temperaturas máxima, mínima y promedio en los Estados Unidos
- Variaciones de temperaturas máximas por año en Australia
- La variación del promedio mundial de temperatura a lo largo de los años
- Premedios de temperaturas máxima en todos los países del mundo a lo largo de los años

Se utilizan saltos de tiempo de 10 años desde 1900 a 2018. Esto porque el dataset es demasiado grande y demoraría mucho ejecutar programas tan pesados en el cluster.

4. Resultados

4.1. Temperatura promedio global por año desde 1900

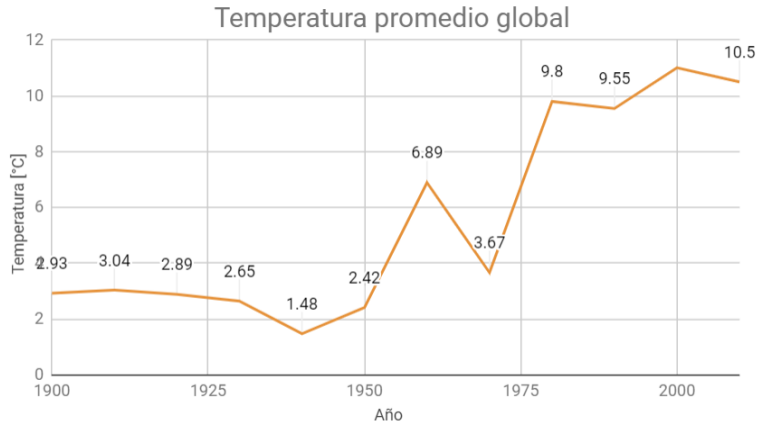

Figura 4: Gráfico que representa la temperatura promedio del mundo con respecto al tiempo

4.2. Temperaturas máximas de Australia de los últimos 118 años

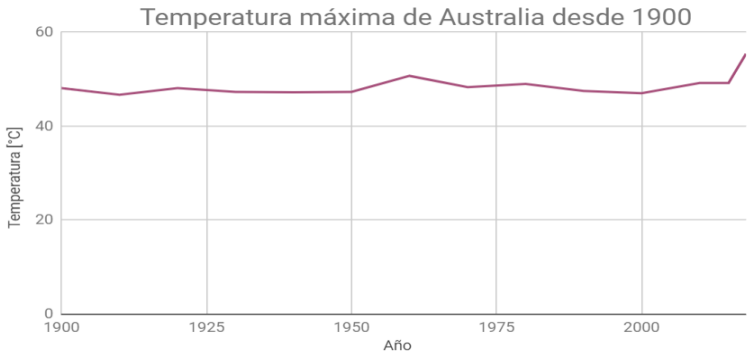

Figura 5: Gráfico que representa la temperatura máxima de Australia con respecto al tiempo

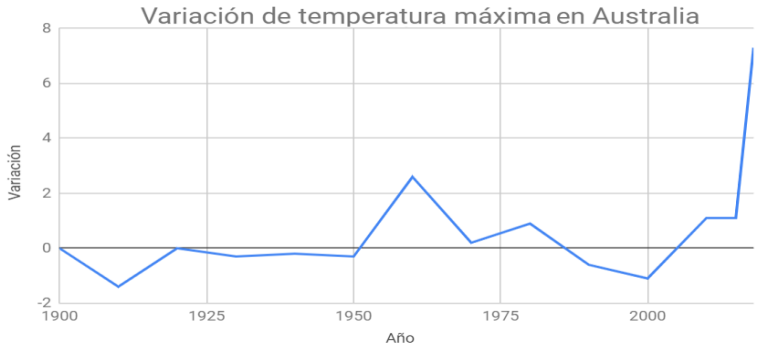

Figura 6: Gráfico que representa la variación de la temperatura máxima de Australia con respecto al tiempo

4.3. Promedio de temperaturas máximas,mínimas y promedio por año en US

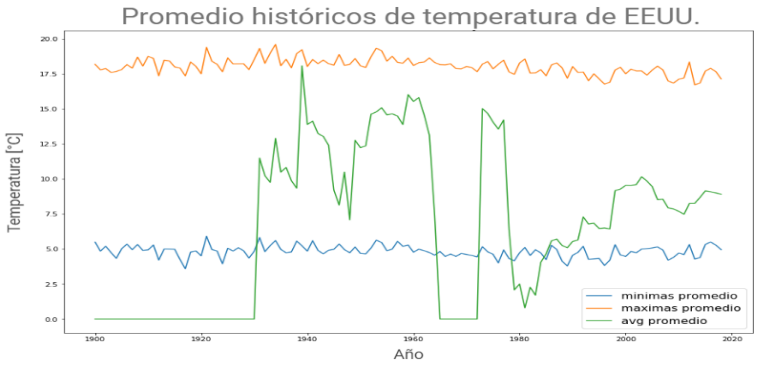

Figura 7: Gráfico que representa el promedio de temperaturas máximas, mínimas y temperatura promedio por año en EEUU.

4.4. Temperaturas promedio de temperaturas máx por país

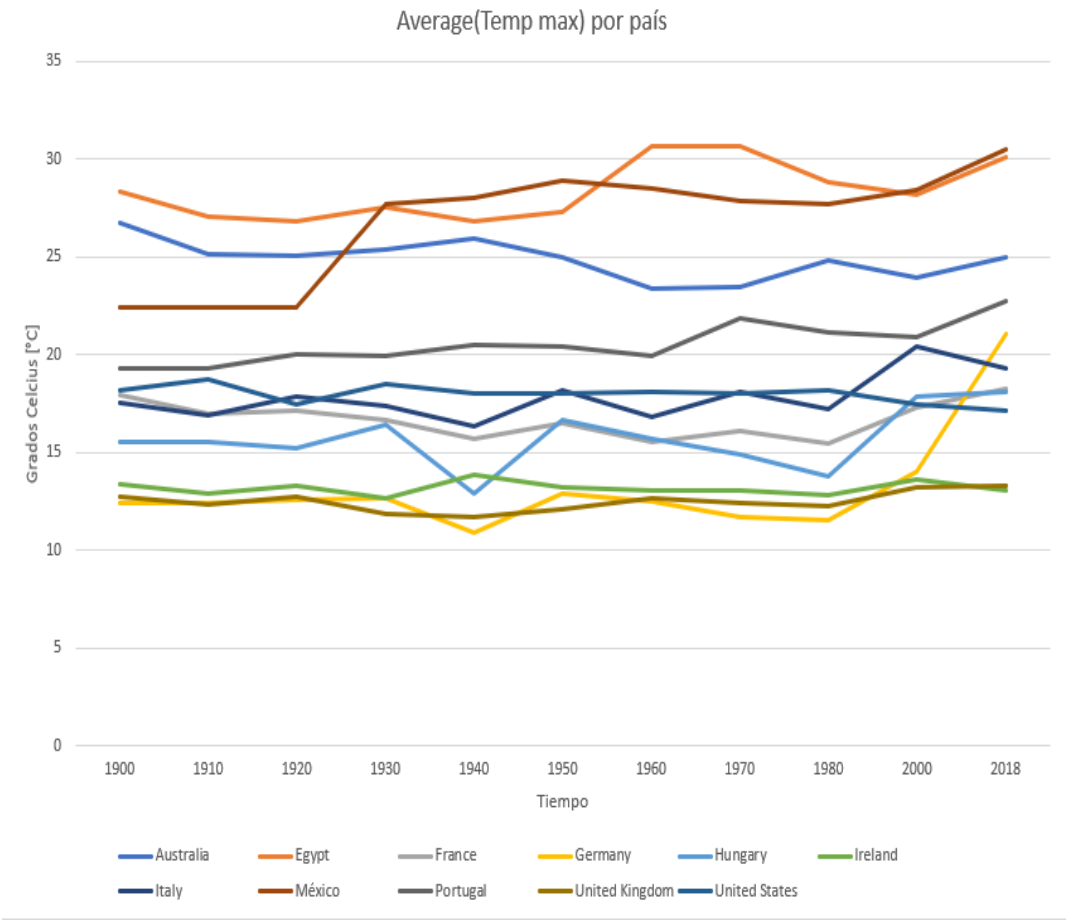

Figura 8: Temperaturas promedio de temperaturas máx por país a lo largo del tiempo

5. Conclusiones y aprendizaje

Pig es una buena herramienta para realizar map/reduce y tiene el beneficio que se parece a la sintáxis de SQL, por lo que se percibe en forma familiar para el usuario.

Respecto de los datos, es posible concluir que existe una tendencia a un alza en la temperatura promedio de los países del mundo y en particular en Australia. Esto es apreciable en las Figuras adjuntas al reporte.

Se puede decir además que los datasets son confiables puesto que provienen de una organización gubernamental. Además los resultados obtenidos calzan con los resultados esperados (búsquedas en la web: papers sobre el tema entre otros), esto valida el dataset aún más.