IDbootcamps

Bootcamp en

# Data Science

Part-Time  - Promoción octubre 2021

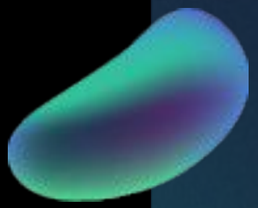
**Lead Instructor:** Marco Russo
**Lead Instructor:** Daniel Montes Serrano

# Data Science Part-Time

*Promoción Octubre 2021*

## 02. Challenge BigQuerython

**marzo 2022**



ID boot camps

# Introducción

Como ya visteis en estos días, dadas las circunstancias tan especiales de centrarnos en la ETL completa desde ingesta, transformación hasta productivizar nuestra aplicación, hemos propuesto realizar este pequeño hackaton/datathon para amenizar un poco estos días.

Lo básico, hay que hacer una serie de cálculos, sencillos, sobre un dataset de datos, y posteriormente, la comparativa será en el tiempo que tarda el script de cada uno de vosotros en ejecutarse en una máquina virtual en Google Cloud.

El proyecto de Google Cloud al que tendréis acceso estará disponible a breve.

IDbootcamps

# Datos

**Tenemos los datos de la natalidad de los EEUU desde 1969 hasta 2010(aproximadamente).**

**¿Por qué EEUU? Porque suelen ser los que más datos tienen disponibles al público.**

**Los datos están subidos a un bucket de Google Cloud Storage en formato csv.**

**Hay 3 tipos de ficheros:**

Natalidad (natalidad*.csv): **contienen la información de la natalidad.**

Raza (race.csv): **contiene las dimensiones de raza.**

Sexo (sex.csv): **contiene las dimensiones de sexo.**

**El esquema de los ficheros lo podéis encontrar aquí:**

https://docs.google.com/spreadsheets/d/1QW9_7eZrUf3BMLgrLqsLPRAfA0BLw5Xna70Xf0V8kYA/edit?usp=sharing

# ¿Cómo vamos a hacer las pruebas?

**Las pruebas se harán en Google Cloud, en una máquina virtual con docker, para facilitar que no haya problemas de en** *mi local funciona*.

**Se ejecutará el siguiente comando para medir el tiempo:**

```
time docker run thebridgebqscript
```

**El script lo ejecutaremos 3 veces y se tomará el mejor tiempo.**

# ¿Cómo vamos a hacer las pruebas?

**La máquina virtual tendrá 2 núcleos, 8gb de ram y un disco de 10gb, para que lo tengáis en cuenta, es la misma que estamos utilizando en Google Colab.**

**Para evitarnos latencias y diferencias a la hora de la ejecución al acceder a los datos, estarán copiados en la carpeta** /data **de la máquina a la que podréis acceder, tenedlo en cuenta a la hora de preparar el docker.**

**Deberéis subir el código a un repositorio público (github) o darnos acceso a donde lo tengáis, así como las instrucciones para generar la imagen de docker para ejecutarlo.**

IDbootcamps

# Salida esperada

## El objetivo es un fichero que contenga la siguiente información:

*Estado (string)*

*B70: Nacimientos en la decada los 70 en ese estado (number)*

*B80: Nacimientos en la decada los 80 en ese estado (number)*

*B90: Nacimientos en la decada los 90 en ese estado (number)*

*B00: Nacimientos en la decada los 2000 en ese estado (number)*

*Race70: Raza con mayor número de nacimientos en la decada de los 70 en ese estado (string)*

*Race80: Raza con mayor número de nacimientos en la decada de los 80 en ese estado (string)*

*Race90: Raza con mayor número de nacimientos en la decada de los 90 en ese estado(string)*

*Race00: Raza con mayor número de nacimientos en la decada de los 2000 en ese estado (string)*

*Male: Numero de nacimientos de hombres en los desde el 70 al 2010 (number)*

*Female: Numero de nacimientos de hombres en los desde el 70 al 2010 (number)*

*Weight: peso medio en kilos de todos los niños nacidos en ese estado desde el 70 al 2010 (float)*

**Formato: .csv**

**El fichero de salida deberéis dejarlo en la carpeta *output* con un nombre identificativo, poned vuestro nombre, id de github o correo de slack por ejemplo, sobre todo para saber quién gana.**

# Presentación de resultados y proclamación de ganador

Una vez tengamos los tiempos, nos pondremos en contacto con los 2 mejores a lo largo de la semana siguiente para que nos hagan una pequeña descripción del código y el por qué de hacerlo con ese lenguaje y librerías (hemos visto que hay un lenguaje que gana por mayoría aplastante, así que esperamos que haya un poco de variedad en las librerías utilizadas, que no hay pocas precisamente ;-) ), no tiene que ser muy grande, unas 10-15 líneas.

Posteriormente se presentarán los resultados y anunciaremos el ganador.

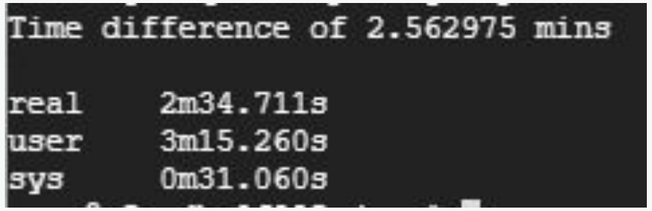
Lo ideal sería hacer un pequeño debate, pero dadas las circunstancias, con que salga un post nos conformamos :-P.

IDbootcamps

# FAQ

- **P: Sería interesante dar tiempos orientativos porque sino no hay "nada que superar" y se va a ciegas. Algo así como entre 5-10 minutos muy bueno, menos de 5 excepcional ....**

**R:** Para que os hagáis una idea, en la máquina que se va a probar, con R ha tardado esto:




- **P: El docker puede hacer uso de todos los recursos de la máquina ?**

**R:** Si, no hay restricciones, los 4 núcleos, los 8GB de ram y los 10GB de ssd (descontando los del sistema y los datos)

# FAQ

- **P: Si no quitamos el arranque de las cuentas (si es representativo en el total) a los lenguajes sobre JVM los perjudica**

**R:** La compilación de docker no se tiene en cuenta, lo que se va a medir es solamente el tiempo de ejecución del tratamiento de datos

- **Se pueden usar herramientas ? o habría que dejar claro que tiene que ser un script ?**

**R:** se puede usar lo que queráis siempre que esté contenido en el docker y se ejecute dentro de este

- **P: ¿El fichero de salida tiene que ser exactamente igual que el ejemplo de la pág 8?**

**R:** Si, el fichero tiene que tener esas columnas para poder comparar unos con otros

- **P:** ¿Se tiene en cuenta el promedio de las 3 ejecuciones?

**R:** No, se toma la más rápida.

# FAQ

- **P: ¿No puedo preprocesar esto y dejarlo preparado en la máquina docker no?**

**R:** No, los datos van a estar en la carpeta /data, el preproceso tiene que ser parte del procesado de datos

- **P: ¿Los registros de natalidad son todos nacimientos exitosos o importa el valor de los campos born_alive_alive, born_alive_dead?**

**R:** no se tiene en cuenta si salió adelante el bebe o no, solamente los nacimientos.

- **P: ¿Los registros de natalidad cuya columna "plurality" es mayor a 1, se debe entender como un nacimiento o como X nacimientos de acuerdo al valor de dicha columna?**

**R:** La columna plurality yo no la he tenido en cuenta para calcular el resultado porque no tenía una descripción

- **P: ¿ Las resultados serán validados? o solo se tendrá en cuenta la rapidez del procesamiento de datos?**

**R:** Sí, se hará una validación de los resultados.

# ¿Preguntas?

Please contact **marco@idbootcamps.com**

Cualquier duda respecto a GCP consultaremos los labs de la semana.