Green Computing en la era de los modelos masivos de IA

PyDay Cali 2024

Autor: Ing. Carlos Andrés Sierra, M.Sc.
carlos.andres.sierra.v@gmail.com

Profesor
Ingeniero de Sistemas
Magister en Ingeniería de Sistemas y Computac ión

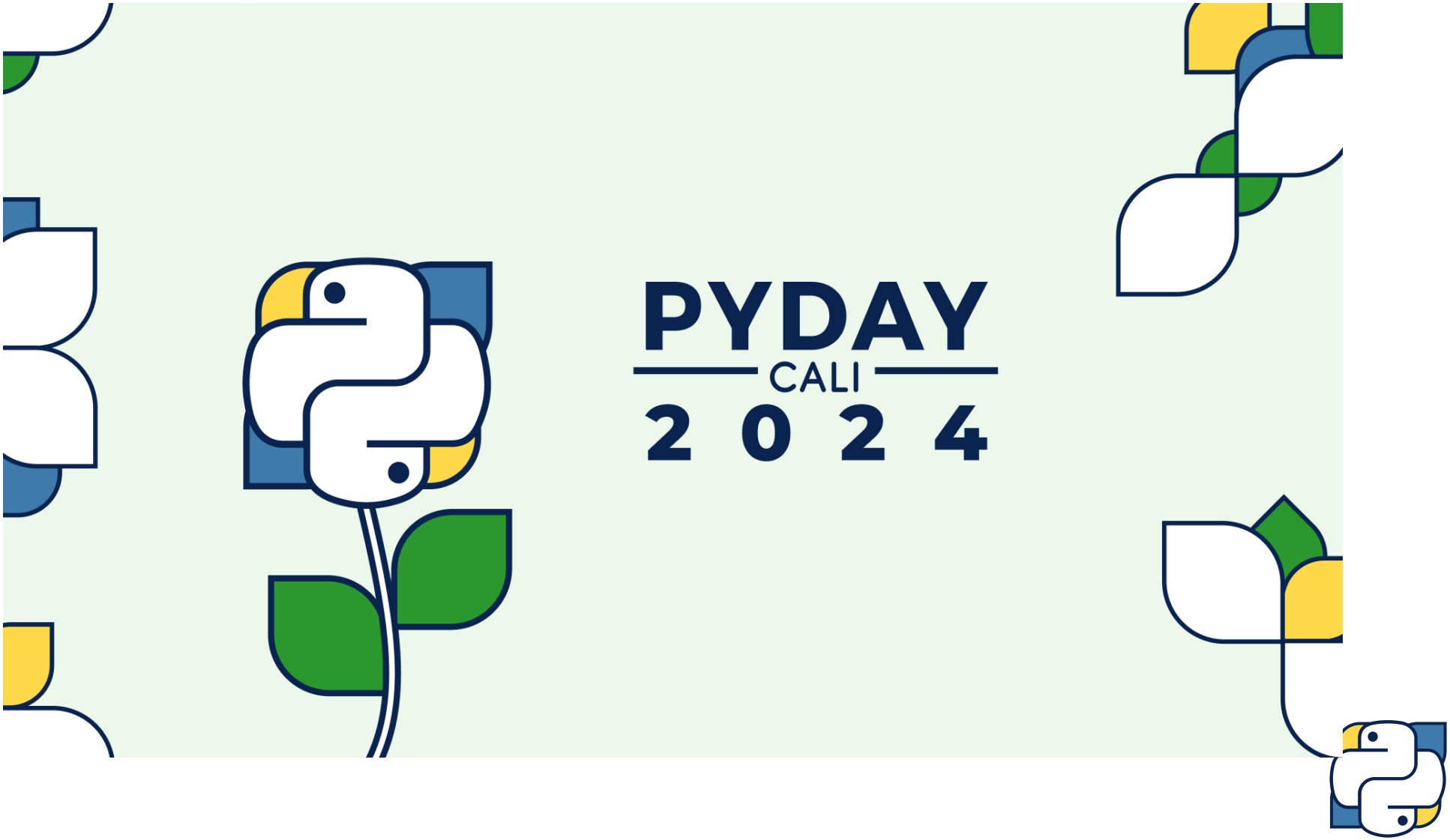
Octubre 19 2024

Outline

Outline

Welcome!

Mucho trabajo, mucho esfuerzo, mucho corazón



Usted no sabe quién soy yo. . . y no lo culpo

- *PyCon Colombia* y *Python Bogotá* **co-organizador**.
- He trabajado como **Ingenierp de Software**, **Cientfico de Datos**, y **Líder Técnico** de Machine Learning.
- **Profesor** en la *Universidad Distrital Francisco José de Caldas* e **Ingeniero de Software/ML/MLOps** en *GenLogs*.
- **Conferencista** en conferencias IEEE, universidades, meetups, . . .

Usted no sabe quién soy yo. . . y no lo culpo

- *PyCon Colombia* y *Python Bogotá* **co-organizador**.
- He trabajado como **Ingenierp de Software**, **Cientfico de Datos**, y **Líder Técnico** de Machine Learning.
- **Profesor** en la *Universidad Distrital Francisco José de Caldas* e **Ingeniero de Software/ML/MLOps** en *GenLogs*.
- **Conferencista** en conferencias IEEE, universidades, meetups, . . .

Usted no sabe quién soy yo. . . y no lo culpo

- *PyCon Colombia* y *Python Bogotá* **co-organizador**.
- He trabajado como **Ingenierp de Software**, **Cientfico de Datos**, y **Líder Técnico** de Machine Learning.
- **Profesor** en la *Universidad Distrital Francisco José de Caldas* e **Ingeniero de Software/ML/MLOps** en *GenLogs*.
- **Conferencista** en conferencias IEEE, universidades, meetups, . . .

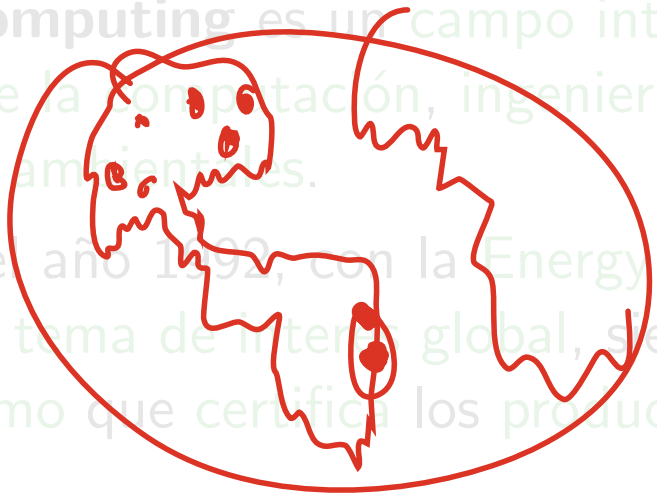
Usted no sabe quién soy yo. . . y no lo culpo

- *PyCon Colombia* y *Python Bogotá* **co-organizador**.
- He trabajado como **Ingenierp de Software**, **Cientfico de Datos**, y **Líder Técnico** de Machine Learning.
- **Profesor** en la *Universidad Distrital Francisco José de Caldas* e **Ingeniero de Software/ML/MLOps** en *GenLogs*.
- **Conferencista** en conferencias IEEE, universidades, meetups, . . .

Outline

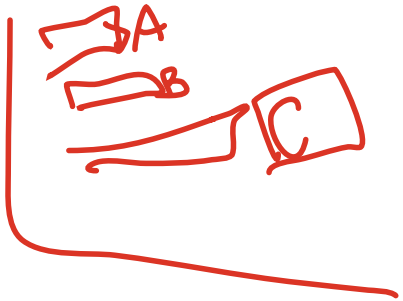
Definiciones Básicas

- **Green Computing** es una disciplina que se enfoca en reducir el consumo de energía y minimizar el impacto ambiental de los sistemas de cómputo.
- **Green Computing** también se conoce como Green IT o Sustainable IT.
- **Green Computing** es un campo interdisciplinario que combina ciencias de la computación, ingeniería eléctrica, ingeniería ambiental, y ciencias ambientales.
- Inicia en el año 1992, con la Energy Star Program de la EPA. Hoy en día, es un tema de interés global, siendo el Green Electronic Council el organismo que certifica los productos electrónicos.



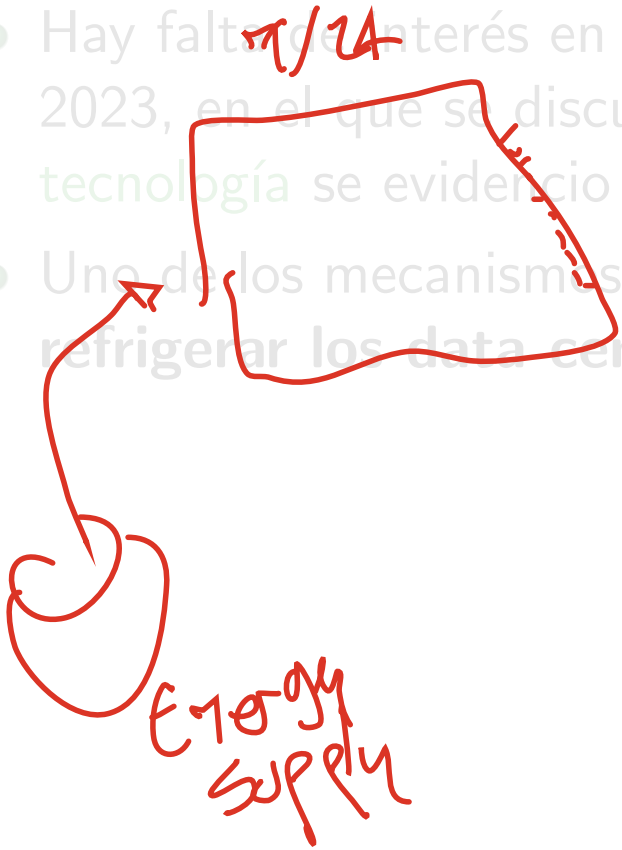
Definiciones Básicas

- **Green Computing** es una disciplina que se enfoca en reducir el consumo de energía y minimizar el impacto ambiental de los sistemas de cómputo.
- **Green Computing** también se conoce como Green IT o Sustainable IT.
- **Green Computing** es un campo interdisciplinario que combina ciencias de la computación, ingeniería eléctrica, ingeniería ambiental, y ciencias ambientales.
- Inicia en el año 1992, con la Energy Star Program de la EPA. Hoy en día, es un tema de interés global, siendo el Green Electronic Council el organismo que certifica los productos electrónicos.



Esfuerzos de la Industria y los Gobiernos

- Un de los mayores retos de la industria y los gobiernos es reducir el consumo de energía de los centros de datos.
- Hay falta de interés en el COP28, realizado en Dubai en Diciembre de 2023, en el que se discutió sobre el impacto ambiental de la tecnología se evidencio esto.
- Uno de los mecanismos más interesantes es la forma en que se pueden **refrigerar los data centers**, como el caso de Facebook en Suecia.

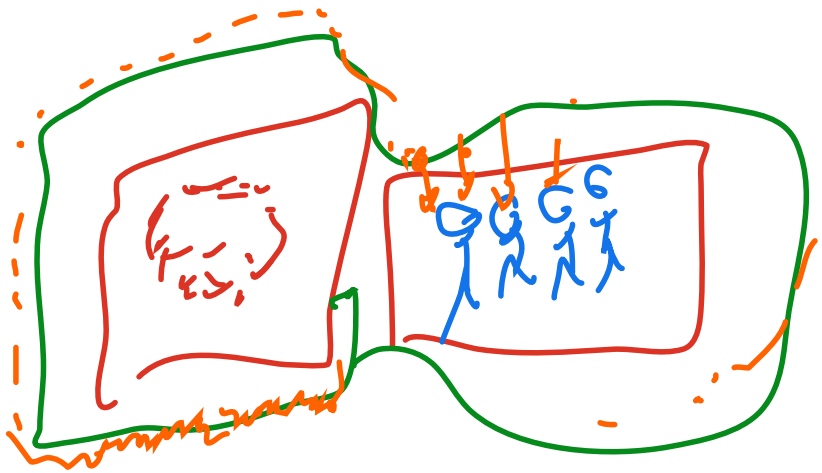


Esfuerzos de la Industria y los Gobiernos

- Un de los mayores retos de la industria y los gobiernos es reducir el consumo de energía de los centros de datos.
- Hay falta de interés en el COP28, realizado en Dubai en Diciembre de 2023, en el que se discutió sobre el impacto ambiental de la tecnología se evidencio esto.
- Uno de los mecanismos más interesantes es la forma en que se pueden **refrigerar los data centers**, como el caso de Facebook en Suecia.

Esfuerzos de la Industria y los Gobiernos

- Un de los mayores retos de la industria y los gobiernos es reducir el consumo de energía de los centros de datos.
- Hay falta de interés en el COP28, realizado en Dubai en Diciembre de 2023, en el que se discutió sobre el impacto ambiental de la tecnología se evidencio esto.
- Uno de los mecanismos más interesantes es la forma en que se pueden **refrigerar los data centers**, como el caso de Facebook en Suecia.



Realidad del Consumo en la Industria de las Telecomunicaciones y la Información

Los gases invernadero son emisiones que contribuyen al calentamiento global.

Realidad del Consumo en la Industria de las Telecomunicaciones y la Información

- La industria de las telecomunicaciones y la información es responsable de una parte significativa de las emisiones de gases invernadero. Se estima que representa entre el 2% y el 3% de las emisiones globales.
- La energía que consumen los centros de datos es responsable de entre el 1% y el 2% de las emisiones globales.
- En cuanto al consumo de energía, se estima que los centros de datos consumen aproximadamente el 1% de la energía eléctrica del mundo. Esto corresponde a aproximanente 200 TWh al año.
- Para dar contexto, un vehículo promedio consume aproximadamente 4 MWh al año, y genera aproximadamente 2 toneladas de CO_2, porcentualmente esto es 0.0002% de las emisiones globales.
- Una vaca produce aproximadamente 100 kg de metano al año, y genera aproximadamente 0.2 toneladas de CO_2, porcentualmente esto es 0.00001% de las emisiones globales.

Realidad del Consumo en la Industria de las Telecomunicaciones y la Información

- La industria de las telecomunicaciones y la información es responsable de una parte significativa de las emisiones de gases invernadero. Se estima que representa entre el 2% y el 3% de las emisiones globales.
- La energía que consumen los centros de datos es responsable de entre el 1% y el 2% de las emisiones globales.
- En cuanto al consumo de energía, se estima que los centros de datos consumen aproximadamente el 1% de la energía eléctrica del mundo. Esto corresponde a aproximanente 200 TWh al año.
- Para dar contexto, un vehículo promedio consume aproximadamente 4 MWh al año, y genera aproximadamente 2 toneladas de CO2, porcentualmente esto es 0.0002% de las emisiones globales.
- Una vaca produce aproximadamente 100 kg de metano al año, y genera aproximadamente 0.2 toneladas de CO2, porcentualmente esto es 0.00001% de las emisiones globales.

Realidad del Consumo en la Industria de las Telecomunicaciones y la Información

- La industria de las telecomunicaciones y la información es responsable de una parte significativa de las emisiones de gases invernadero. Se estima que representa entre el 2% y el 3% de las emisiones globales.
- La energía que consumen los centros de datos es responsable de entre el 1% y el 2% de las emisiones globales.
- En cuanto al consumo de energía, se estima que los centros de datos consumen aproximadamente el 1% de la energía eléctrica del mundo. Esto corresponde a aproximanente 200 TWh al año.
- Para dar contexto, un vehículo promedio consume aproximadamente 4 MWh al año, y genera aproximadamente 2 toneladas de CO2, porcentualmente esto es 0.0002% de las emisiones globales.
- Una vaca produce aproximadamente 100 kg de metano al año, y genera aproximadamente 0.2 toneladas de CO2, porcentualmente esto es 0.00001% de las emisiones globales.

Realidad del Consumo en la Industria de las Telecomunicaciones y la Información

- La industria de las telecomunicaciones y la información es responsable de una parte significativa de las emisiones de gases invernadero. Se estima que representa entre el 2% y el 3% de las emisiones globales.
- La energía que consumen los centros de datos es responsable de entre el 1% y el 2% de las emisiones globales.
- En cuanto al consumo de energía, se estima que los centros de datos consumen aproximadamente el 1% de la energía eléctrica del mundo. Esto corresponde a aproximanente 200 TWh al año.
- Para dar contexto, un vehículo promedio consume aproximadamente 4 MWh al año, y genera aproximadamente 2 toneladas de CO2, porcentualmente esto es 0.0002% de las emisiones globales.
- Una vaca produce aproximadamente 100 kg de metano al año, y genera aproximadamente 0.2 toneladas de CO2, porcentualmente esto es 0.00001% de las emisiones globales.

Punto de vista de NVIDIA

Transitioning all AI and HPC on CPU Servers to GPUs

- *10TWh AI and 1TWh HPC savings from transitioning CPU servers to GPU servers (assume AI CPU-servers are currently running inference)*
- *1.8M AI servers in use and 145K HPC servers in use*
- *Assumption: Few CPU servers are used for AI-Training*
- *Additional savings from switching Graphics, Data Analytics workloads*

Offloading CPU Operations to DPUs

- *AWS estimates 30% of CPU cores are required for Hypervisor and Operational Management overhead that can be offloaded to the DPU*
- *25% of servers currently DPU accelerated based on Crehan estimates*
- *2022 DPU Server Shipment TAM is 14.2 M servers (25% accelerated = 3.55M)*
- *Each DPU adds only 75 Watts*
- *The increased performance with DPU results in a reduction of 1.065M servers to run the same amount of workloads*
- *Resulting in total available power savings of 7.7 TWh*

Que envidia a la NVIDIA

Buenas Prácticas

Para un usuario normal se recomienda:

- Disminuir el `brillo de la pantalla`, esto se estima puede ahorrar hasta un 20% de `energía`.
- Usar modos de ahorro de energía, o configurar el sistema para que se apague la pantalla después de un tiempo de inactividad.
- Utilizar electrónicos con certificaciones de eficiencia energéticas, como **Energy Star**.
- Desconectar los dispositivos electrónicos cuando no se están utilizando.
- Borrar los **emails no requeridos**. Se estima que cada email guardado de forma indefinida consume 0.3 gramos de CO2 al año.

Outline

Modelos de IA actuales

- Se estima que OpenAI ha gastado aproximadamente 1000 TWh en entrenar el modelo ChatGPT.
- Esto es aproximadamente el 0.5% de la energía eléctrica del mundo.
- GPT-4 se estima requirio entre 20000 y 30000 unidades de procesamiento, durante más de 3 semanas, y un costo de 190 **millones de dólares**.

Modelos de IA actuales

- Se estima que OpenAI ha gastado aproximadamente 1000 TWh en entrenar el modelo ChatGPT.
- Esto es aproximadamente el 0.5% de la energía eléctrica del mundo.
- GPT-4 se estima requirio entre 20000 y 30000 unidades de procesamiento, durante más de 3 semanas, y un costo de 190 **millones de dólares**.

Green Coding

- Green coding es una **práctica** que se enfoca en `escribir código` de manera eficiente y sostenible.
- La idea es **minimizar** el consumo de recursos y energía de los programas. Acá entran conceptos como `algoritmos eficientes`, `estructuras de datos eficientes`, y `optimización de código`.
- Pensar en `complejidad algorítmica` y `eficiencia de código` es fundamental para escribir código verde.
- **Python** es un **lenguaje de programación** que *facilita* la escritura de código verde.

Green Coding con Python

- En el mundo de ML, el desarrollo de librerías como TensorFlow y PyTorch ha permitido que los desarrolladores puedan entrenar y desplegar modelos de manera más eficiente, intentando sacar provecho de GPUs y TPUs.
- La necesidad de vivir en espacios numéricos hace que librerías como NumPy sean fundamentales para el desarrollo de código eficiente, usando conceptos como **vectorización**.
- Otra práctica que se puede usar es el profile de código, para identificar cuellos de botella y optimizar el código. En python se puede usar el módulo cProfile.
- El lenguaje de programación C es más eficiente que Python, por lo que se puede usar Cython para compilar código Python a C.

Green Coding con Python

- En el mundo de ML, el desarrollo de librerías como TensorFlow y PyTorch ha permitido que los desarrolladores puedan entrenar y desplegar modelos de manera más eficiente, intentando sacar provecho de GPUs y TPUs.
- La necesidad de vivir en espacios numéricos hace que librerías como NumPy sean fundamentales para el desarrollo de código eficiente, usando conceptos como **vectorización**.
- Otra práctica que se puede usar es el profile de código, para identificar cuellos de botella y optimizar el código. En python se puede usar el módulo cProfile.
- El lenguaje de programación C es más eficiente que Python, por lo que se puede usar Cython para compilar código Python a C.

Green Coding con Python

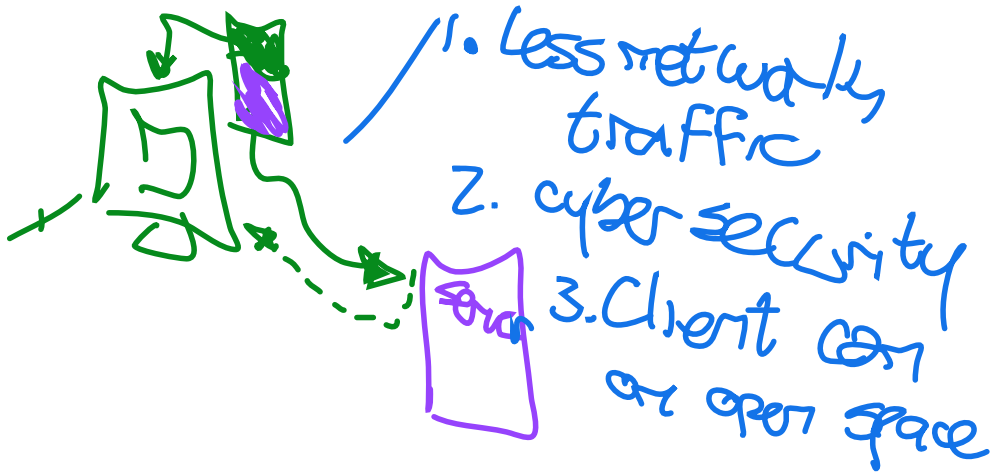
- En el mundo de ML, el desarrollo de librerías como TensorFlow y PyTorch ha permitido que los desarrolladores puedan entrenar y desplegar modelos de manera más eficiente, intentando sacar provecho de GPUs y TPUs.
- La necesidad de vivir en espacios numéricos hace que librerías como NumPy sean fundamentales para el desarrollo de código eficiente, usando conceptos como **vectorización**.
- Otra práctica que se puede usar es el profile de código, para identificar cuellos de botella y optimizar el código. En python se puede usar el módulo cProfile.
- El lenguaje de programación C es más eficiente que Python, por lo que se puede usar Cython para compilar código Python a C.

Green Coding con Python

- En el mundo de ML, el desarrollo de librerías como TensorFlow y PyTorch ha permitido que los desarrolladores puedan entrenar y desplegar modelos de manera más eficiente, intentando sacar provecho de GPUs y TPUs.
- La necesidad de vivir en espacios numéricos hace que librerías como NumPy sean fundamentales para el desarrollo de código eficiente, usando conceptos como **vectorización**.
- Otra práctica que se puede usar es el profile de código, para identificar cuellos de botella y optimizar el código. En python se puede usar el módulo cProfile.
- El lenguaje de programación C es más eficiente que Python, por lo que se puede usar Cython para compilar código Python a C.

Edge-AI con Python

- Edge-AI es una **técnica** que permite `ejecutar` modelos de *IA* en dispositivos embebidos.
- Edge-AI permite **reducir** el consumo de energía y *mejorar* la `privacidad` de los *usuarios*.
- Con Python se tienen **librerías** como TensorFlow Lite y PyTorch Mobile que `permiten entrenar y desplegar` modelos de *IA* en dispositivos embebidos.



Outline


django girls
2024
Barranquilla
Evento gratuito y presencial
INSCRIPCIONES ABIERTAS
Inscríbete: https://bit.ly/dgb24
Sábado 23 de Noviembre
2024
8:30 a 16:30
Universidad del Norte


django girls
Barranquilla
2024
Evento gratuito y presencial
BUSCAMOS
GUÍAS
https://bit.ly/dgb24
Sábado 23 de Noviembre 2024
8:30 a 16:30
Universidad del Norte

Gracias!!

Linkedin: *casierrav*