Tecnologías emergentes: blockchain, aprendizaje federado

Grado en Ingeniería Informática
Universidad de Burgos

UNIVERSIDAD DE BURGOS

José Ignacio Santos, José Manuel Galán
jisantos @ ubu.es, jmgalan @ ubu.es

Contenidos

- Blockchain
  - Cómo funciona
  - SHA
  - Blockchain distribuido
  - Minado
  - Tokens
  - Utilización como libro mayor
  - Aplicaciones
- Aprendizaje federado
  - Arquitectura
  - Dificultades
  - Aplicaciones

Referencias

- Anders Brownworth, live blockchain (MIT) http://blockchain.mit.edu/how-blockchain-works
- Anders Brownworth, Blockchain demo https://andersbrownworth.com/blockchain/hash
- Blockchain business models, Duke University (Coursera)
- Xu, Jie, Benjamin S. Glicksberg, Chang Su, Peter Walker, Jiang Bian, y Fei Wang. «Federated Learning for Healthcare Informatics». Journal of Healthcare Informatics Research 5, n.º 1 (1 de marzo de 2021): 1-19

Motivación

En este tema se introducen tecnología que **potencialmente** pueden generar **grandes cambios** en productos/servicios, modelos de negocio, desde el ámbito de la gestión de la información

En algunas teorías de innovación a este tipo de tecnologías se las llama **tecnología disruptiva** (Christensen, 1995)

Como en todo aquello relacionado con la predicción de los cambios tecnológicos y sus consecuencias en la economía y la sociedad existe una gran volatilidad de tecnologías candidatas a ser “disruptivas” y blockchain es una de ellas. En este tema no se hacen predicciones de los cambios que blockchain generará, simplemente describimos sus características principales.

Clayton M. Christensen “Disruptive Technologies: Catching the Wave” 1995

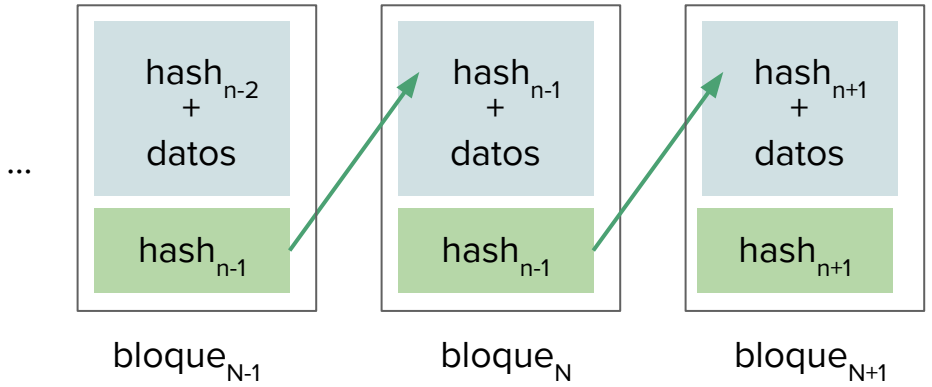
Blockchain

Photo by Jeremy Zero on Unsplash

Definición de blockchain

Blockchain es una **base de datos** construida mediante **bloques enlazados a través de claves criptográficas (hash)**

- Cada bloque tiene su propio hash (resumen mediante una huella digital)
- Dentro del contenido de un bloque se encuentra el hash del bloque predecesor



Este diseño garantiza (=hace muy difícil) la **inmutabilidad** de los datos de los bloques:

- Generar un hash lleva tiempo (**minado**)
- Cambiar un dato supone cambiar los hash del bloque que lo contiene y de todos los bloques sucesivos

¿Cómo funciona blockchain?

Blockchain demo
https://andersbrownworth.com/blockchain/hash

Secure Hash Algorithm (SHA)

¿Cómo generar un hash de un conjunto de datos (mensaje)?
Mediante una **función criptográfica basada en curvas elípticas** que genera una clave digital (e.g. SHA-256 tiene una salida de 256 bits) de un argumento de entrada

- **One-way function**: es fácil generar una firma pero es muy muy difícil obtener el argumento a partir de la firma
- **Lost of information**: la firma digital contiene menos información que el argumento de entrada
- **Algoritmo determinista** (la probabilidad de colisión es muy muy pequeña)

https://en.wikipedia.org/wiki/Elliptic_Curve_Digital_Signature_Algorithm

Ejemplo

Suponer que conozco a una persona de este grupo que ha cometido un delito “académico” y quiero chantajearle.

Publico el siguiente mensaje en redes sociales a través de un perfil falso:

> *Sé quién ha cometido un grave delito académico en la clase de Gestión de la Información del Grado en Ingeniería Informática de la UBU. Si esa persona no envía una transferencia en bitcoins por valor de 1000 euros a la cuenta de Samourai Wallet 123456789 haré pública su identidad. El SHA256 del nombre y apellidos de la persona es:*
>
> *4d130effa2df045685d7c450931393a8abeaab11f9f3a5ea17891e34c6bb065a*

Blockchain distribuido

Blockchain puede implementarse como:

- En un sistema **privado** centralizado/descentralizado, e.g. www.hyperledger.org
- En un sistema **público descentralizado**, e.g. bitcoin, ethereum, ...
  - No existe una autoridad centralizada que tome decisiones, sino que cualquiera puede incorporarse y participar en la red
  - Todos los nodos guardan una copia (redundancia), participan en la verificación e incorporación de bloques (**algoritmos de consenso**)

Un **sistema distribuido refuerza la inmutabilidad** de la base de datos:

- A la dificultad de modificar un bloque-hash y todos los bloques-hash sucesivos se añade la de **difundir el cambio en más del 50% de los nodos** (51% attack)

Minado

Al contenido de un bloque (mensaje + hash predecesor) se le añade un número (**NONCE**)

No vale cualquier número, se impone una dificultad tal que el bloque con el nonce genere un hash con **un número inicial de ceros** (e.g. 14 en bitcoin)

La búsqueda es por fuerza bruta, la dificultad determina el tiempo medio (e.g. 10 minutos en bitcoin)

En **blockchain distribuidos** la creación de bloques y su verificación la hacen los nodos

Los nodos toman un conjunto de transacciones para construir un bloque e intentan hallar un nonce que verifique la dificultad requerida (=**minado**)

Cuando un nodo encuentra un nonce lo publica en la red, el resto de nodos lo verifican y finalmente se incorpora a la cadena de bloques (quedando confirmadas todas las transacciones del nuevo bloque).

El nodo “minero” es **recompensado** (e.g. con tokens de una determinada moneda); es un incentivo para participar y mantener la red de blockchain

MINADO

Proof of Work

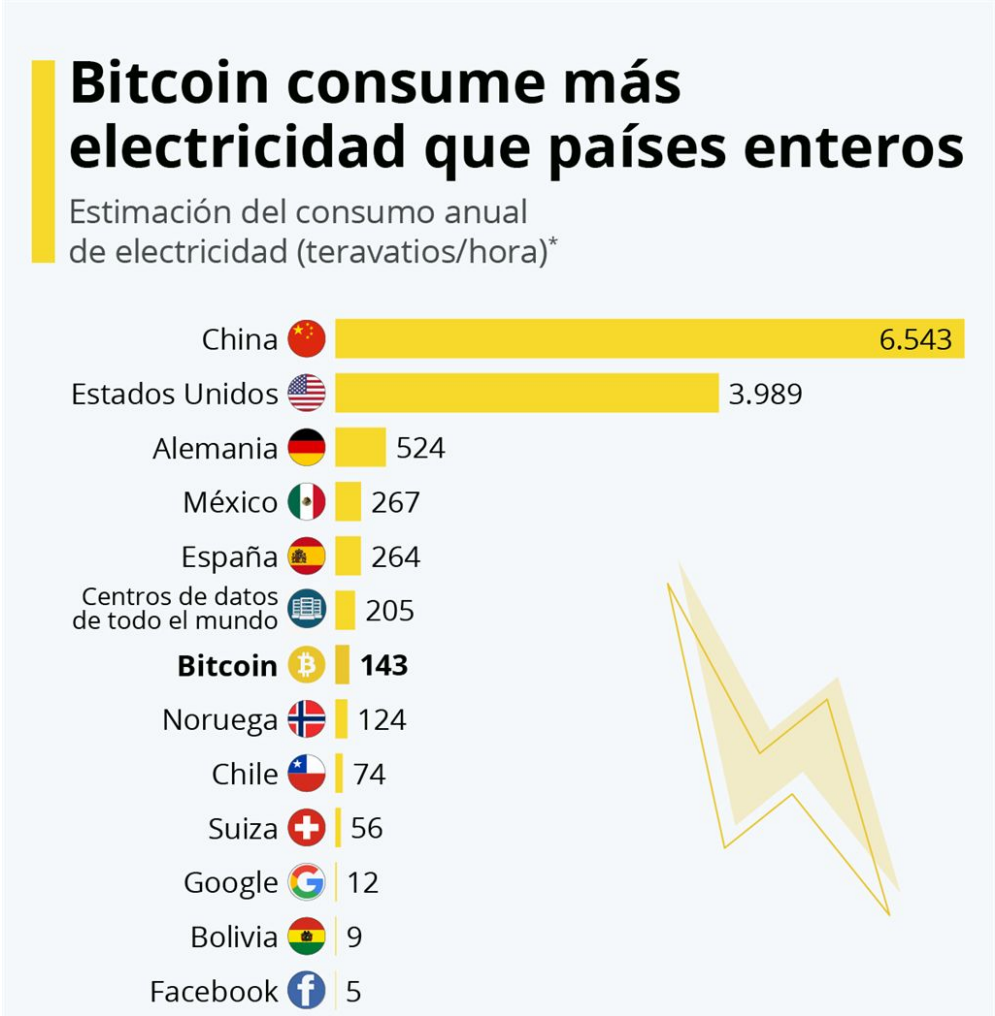
El sistema de minado por el que se exige un **coste “moderadamente caro pero factible”** (e.g. tiempo de computación para calcular el nuevo NOUNCE del bloque) a cada cliente cuya verificación es significativamente más sencilla se denomina “**Proof of Work System**”

Este sistema aumenta la robustez de la base de datos

Sin embargo, requiere de un **consumo energético muy grande** para su funcionamiento:
Cambridge Bitcoin Electricity Consumption Index (CBEC)

Fuente: CBEC

Consumo de Bitcoin



Fuente: Statista

Libro mayor

La estructura de un blockchain es ideal para almacenar un **libro mayor** (**ledger**)

Un libro mayor almacena operaciones contables (un activo se sustrae de una cuenta para añadirse en otra) que representan **transacciones** entre agentes

10/10/2021 Ana (-3€) Rubén (+3€)

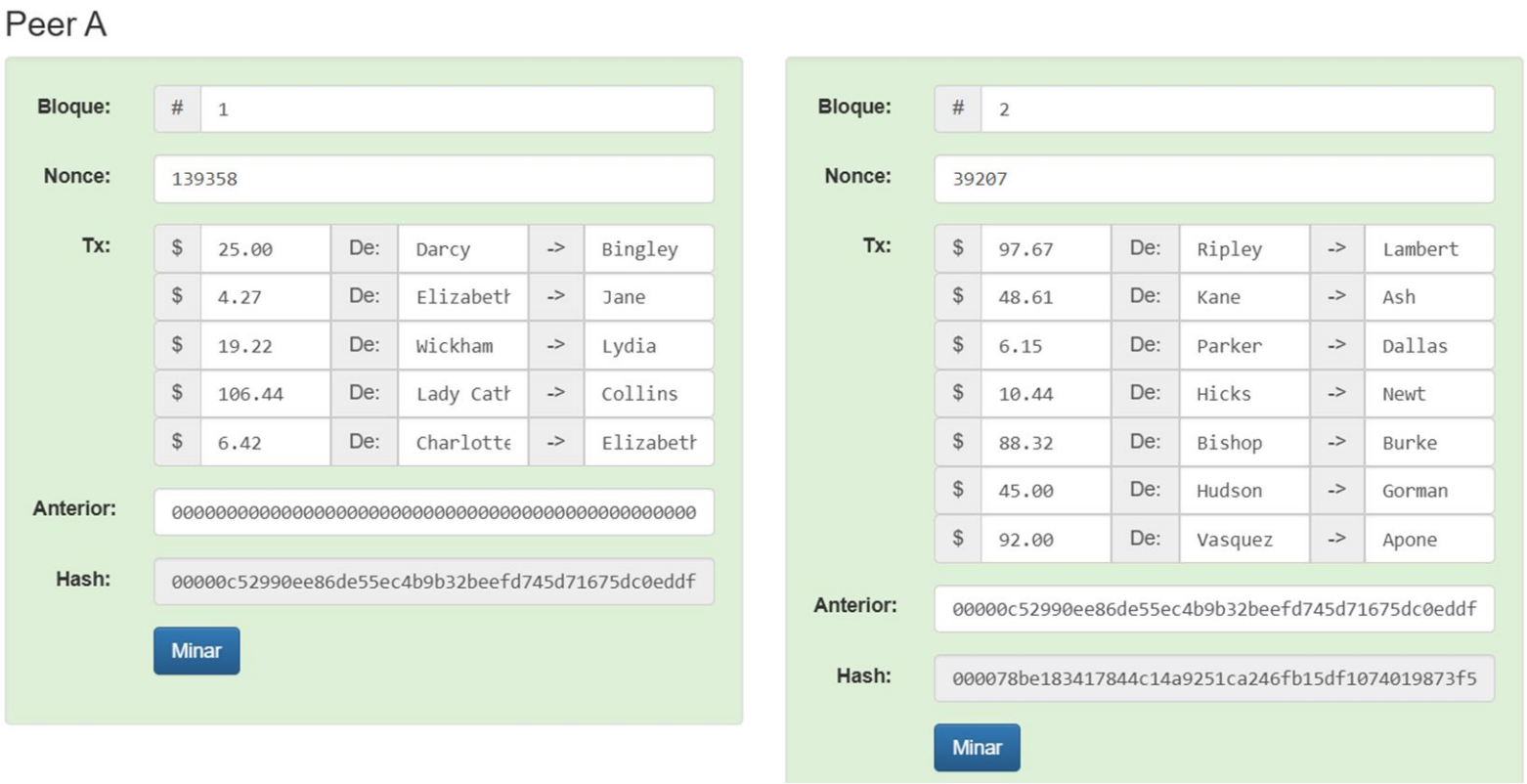
Si las transacciones han sido verificadas (Ana es dueña de los 3 euros) la incorporación en un bloque es garantía de **permanencia sin cambios**, en cualquier momento puedo conocer el estado de un agente, el propietario de un activo, ...

Blockchain “garantiza” la inmutabilidad del libro mayor. La **inmutabilidad está garantiza por**

(1) la estructura de bloques con hash criptográficos,

(2) si además es descentralizado, el mecanismo de consenso necesario para verificar y añadir transacciones

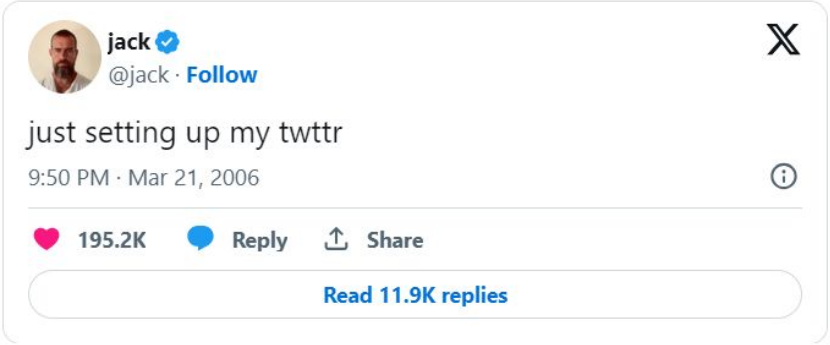
Libro mayor


Peer A
Bloque:
1
Nonce:
139358
Tx:
$ 25.00 De: Darcy -> Bingley
$ 4.27 De: Elizabeth -> Jane
$ 19.22 De: Wickham -> Lydia
$ 106.44 De: Lady Cath -> Collins
$ 6.42 De: Charlotte -> Elizabeth
Anterior:
0000000000000000000000000000000000000000000000000000
Hash:
00000c52990ee86de55ec4b9b32beefd745d71675dc0eddf
Minar
Bloque:
2
Nonce:
39207
Tx:
$ 97.67 De: Ripley -> Lambert
$ 48.61 De: Kane -> Ash
$ 6.15 De: Parker -> Dallas
$ 10.44 De: Hicks -> Newt
$ 88.32 De: Bishop -> Burke
$ 45.00 De: Hudson -> Gorman
$ 92.00 De: Vasquez -> Apone
Anterior:
00000c52990ee86de55ec4b9b32beefd745d71675dc0eddf
Hash:
000078be183417844c14a9251ca246fb15df1074019873f5
Minar

Tokens

Un token es una representación digital única de ALGO:

- Permisos para hacer, entrar, etc.: **UTILITY TOKENS**
- Derechos, participaciones en, valor fiduciario (e.g., moneda), etc.: **SECURITY TOKENS**
- Una obra digital (e.g., foto, dibujo, canción, vídeo,...): **NON FUNGIBLE TOKENS (NFT)**



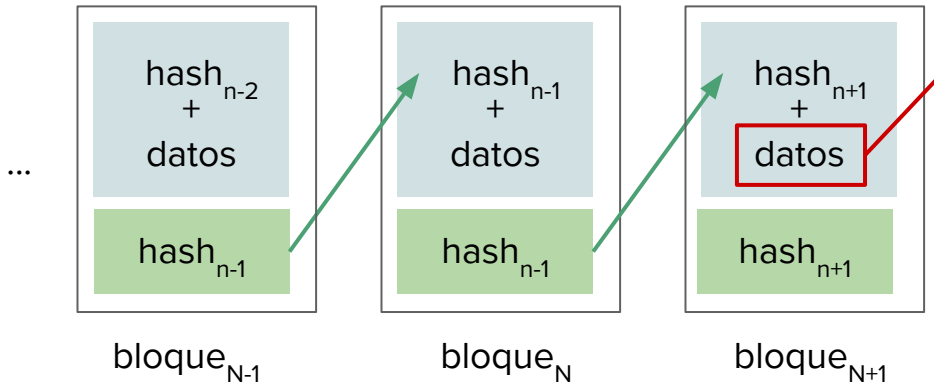
Subastado en 2021 por $2.9M

Los tokens están siempre ligados a un blockchain que los crea y registra las transacciones

Ejemplo: Contratos Inteligentes

Un **contrato inteligente (Smart Contracts)** es un código almacenado en un bloque de blockchain

El código describe **un programa que ejecuta un contrato** entre partes, si se cumplen determinadas condiciones entonces ejecuta una determinada decisión



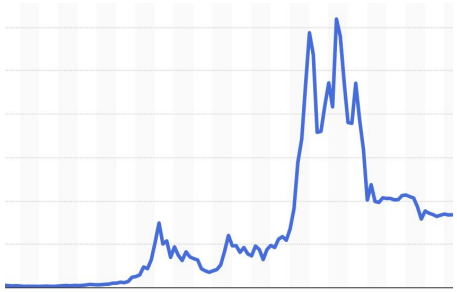

```
pragma solidity >=0.4.22 <0.6.0;

/// @title Voting with delegation.
contract Ballot {
    // This declares a new complex type which will
    // be used for variables later.
    // It will represent a single voter.
    struct Voter {
        uint weight; // weight is accumulated by delegation
        bool voted;  // if true, that person already voted
        address delegate; // person delegated to
        uint vote;   // index of the voted proposal
    }

    // This is a type for a single proposal.
    struct Proposal {
        bytes32 name;   // short name (up to 32 bytes)
        uint voteCount; // number of accumulated votes
    }

    address public chairperson;
```

Tulip Mania

Foto de Pierre Borthiry - Peiobty en Unsplash

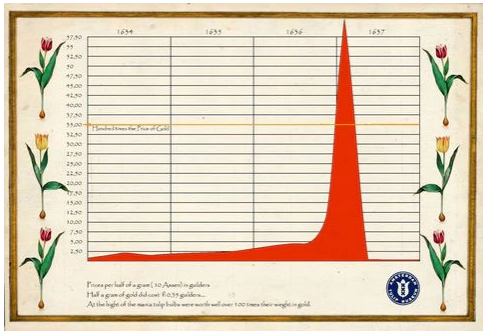
https://youtu.be/OCSxJcX8VDk

Foto de Kwang Mathurosemontri en Unsplash



Bitcoin price 2013-2022 (Statista)

Burbuja de tulipanes (1637): https://en.wikipedia.org/wiki/Tulip_mania



Amsterdam Tulip Museum

BBD tradicionales vs blockchain

Bases de datos relacionales	Blockchain
Estructura de datos en tablas relacionales	Estructura en bloques secuenciales ligados por un hash
Gestión centralizada	Gestión distribuida
“Fácilmente” modificable	Inmutable
Más eficiente y rápida en accesos y consultas	Elevado consumo (proof of work) y lenta en consultas
Aplicaciones generales	Aplicaciones donde la inmutabilidad es clave y los datos tienen dimensión longitudinal (libro mayor)

Resumen de características

Otra forma de ver blockchain es como un **libro mayor distribuido**:

- Conjunto de transacciones temporalmente ordenadas,
- Almacenados en bloques enlazados por un hash criptográfico (huella digital) con un patrón predefinido (que requiere un NOUNCE de complejidad suficiente)
- Distribuido (base de datos completa) en una red de computadores (nodos)

Y esto proporciona las siguientes características a los datos almacenados:

- Los datos son **inmutables** (en la medida en que cambiar un dato exige: (1) cambiar los hash de todos los bloques posteriores, y (2) cambiar las copias de la mayoría de nodos (mecanismo de consenso)
- Los datos **no pueden eliminarse** (cualquier cambio exige una nueva transacción) y son fácilmente **trazables** (desde su origen hasta el momento actual)

Aplicaciones de blockchain

Campbell R. Harvey (2019):

- Votaciones
- Internet of things (IoT)
- Recetas de fármacos
- Estados financieros de las empresas
- Criptomonedas
- Historia Clínica

Blockchain permite gestionar la “**memoria**” de eventos, transacciones, intercambios, decisiones, etc. de forma segura, ágil, y sin necesidad de intermediarios que garanticen la veracidad de los cambios

Ejemplo: NFTS

Un **NFT (Non-Fungible Token)** es un identificador digital que no puede ser copiado o alterado que se almacena asociado a un propietario en un bloque de blockchain

Pueden representar un **activo real o virtual** (cualquier archivo digital, foto, vídeo, música, ...)

El propietario puede **transferir** (transacción almacenada en blockchain) el token

Fuente: Wikipedia

Federated Learning

Ejemplo de infrautilización de los datos clínicos

Uno de los **principales retos** en la gestión de la información en muchos sistemas:

Los **datos están infrautilizados porque están dispersos en diferentes** instituciones, nacionales e internacionales, y las preocupaciones por la privacidad limitan su acceso (Rieke et al, 2020)

- Por ejemplo, imagina el entrenamiento de una IA para la detección mutaciones genéticas asociadas a un mayor riesgo de cáncer ¿cómo recogemos los datos de entrenamiento? ¿un conjunto de hospitales? ¿todo un país? ¿toda la UE?
- Idealmente se podrían compartir los datos, pero existen importantes (y necesarios) requerimientos legales que dificultan/impiden un entrenamiento clásico

Aprendizaje federado

Una solución es el **aprendizaje federado** (AF) que puede definirse como un paradigma de Machine Learning en el que múltiples entidades colaboran para entrenar un modelo compartido manteniendo los datos de entrenamiento localmente sin compartir

El AF es útil cuando existan problemas de:

1. **Privacidad y seguridad en los datos:** AF solo comparte el modelo
2. **Soberanía de los datos**: los datos no salen del ámbito local, respetando las normas y leyes locales

El AF es interesante en diversos ámbitos, por ejemplo en la gestión de la información clínica en la que diferentes entidades (hospitales, centros de investigación, institutos de salud pública, etc.) pueden beneficiarse de mejores modelos garantizando los requerimientos de privacidad, seguridad, y soberanía de los datos clínicos de los pacientes

Arquitectura del aprendizaje federado

Se propone un **modelo global** (mismos parámetros para todos los nodos)

Se define una función de pérdida global, combinación ponderada de las pérdidas del modelo en los datos locales de cada nodo.

$$F(w) = \sum_{k=1}^{K} \frac{n_k}{n} F_k(w)$$

Algoritmo **Federated Average (FedAvg)**:

1) El servidor central envía a los clientes los pesos actuales wt

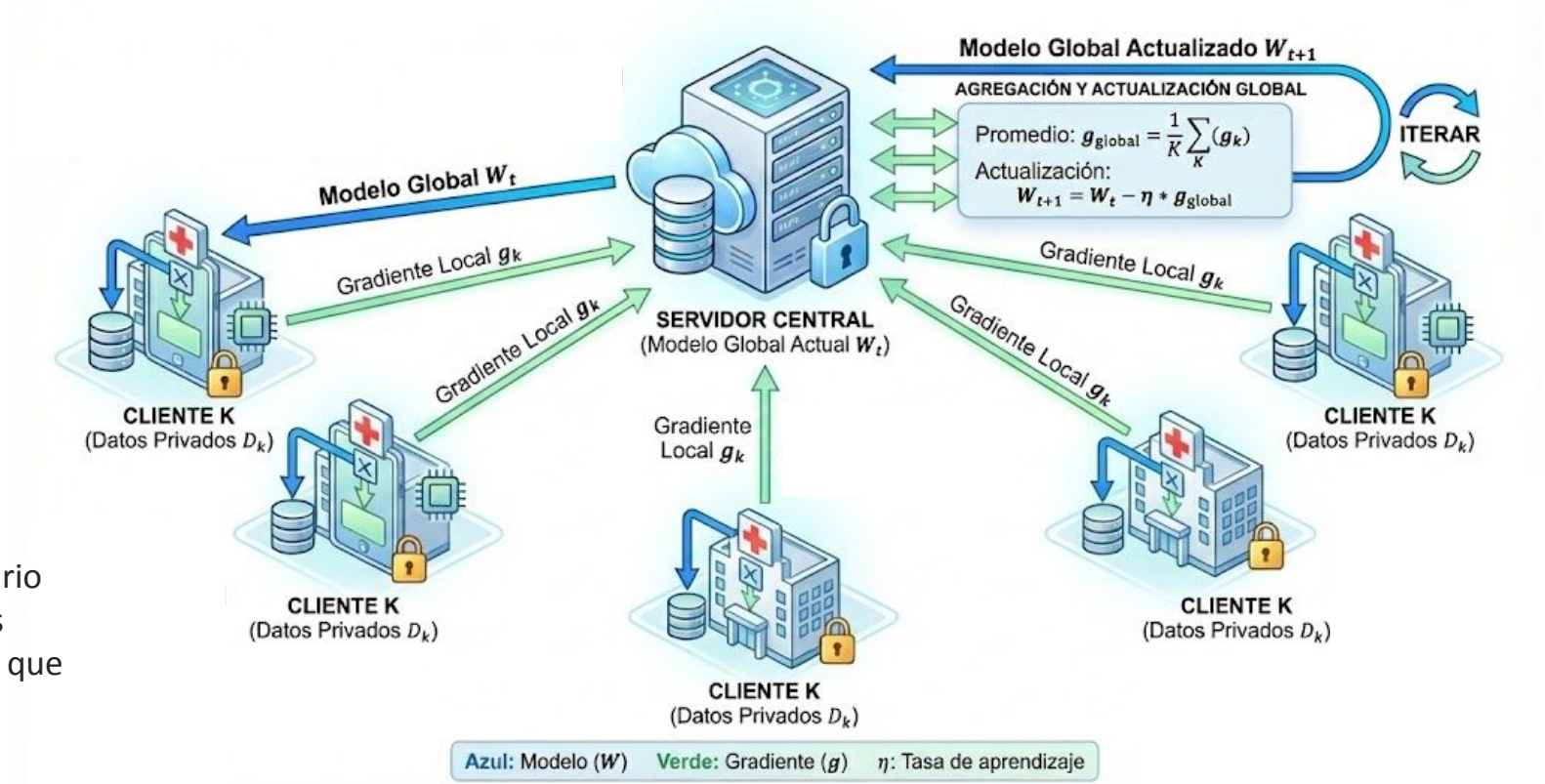
2) Cada cliente ejecuta un SGD sobre sus propios datos durante varias epochs y envía la diferencia

$$\Delta w^k = w^k_{final} - w_t$$

3)El servidor calcula el nuevo modelo global

$$w_{t+1} = w_t + \sum_{k=1}^{K} \frac{n_k}{n} (\Delta w^k)$$

Federated Average (FedAvg)



No es necesario compartir los datos locales que permanecen privados

Dificultades del aprendizaje federado

Algunas de las dificultades de la implementación de modelos de AF son:

- **Heterogeneidad en los datos**: distribuciones muy sesgadas con datos de nodos no representativos del conjunto
- **Escalabilidad**: problemas de rendimiento al aumentar la complejidad del modelo, los tamaños de los datos de entrenamiento o el número de nodos participantes
- **Seguridad y privacidad**: en ocasiones se puede exigir cifrado de los datos para evitar quiebras de seguridad en el proceso de entrenamiento
- **Evaluación del grado de participación** de los nodos: si existen determinantes económicos puede ser necesario determinar la contribución de cada nodo para asignar costes o beneficios

Ejemplos

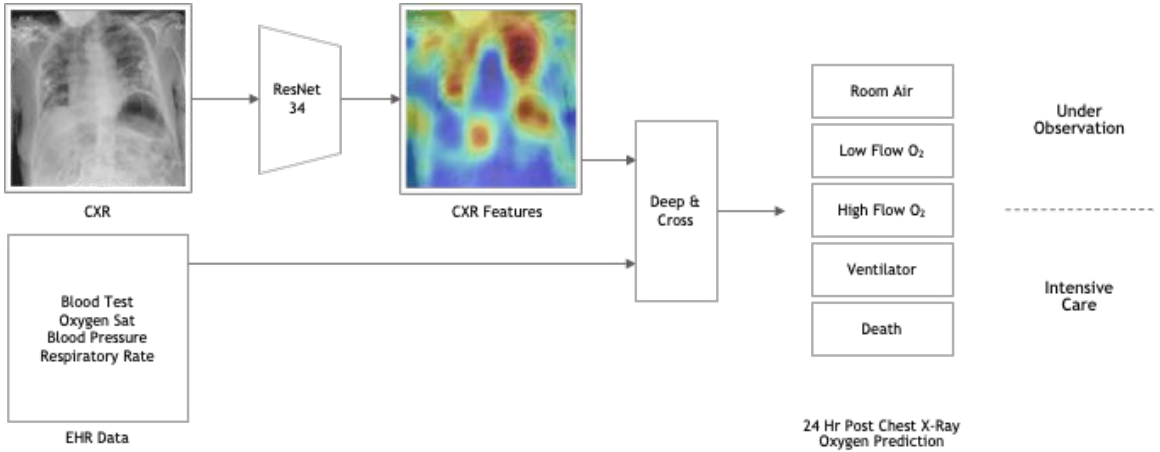
“20 Hospitals in 20 Days Build AI Model that Predicts Oxygen Needs”
https://blogs.nvidia.com/blog/2020/10/05/federated-learning-covid-oxygen-needs/

The **model EXAM** is designed to **predict oxygen requirements in COVID-19 patients**, useful prior to PCR results, using objective, standardized imaging inputs

The global FL model, EXAM, outperformed local models

Participating in FL provides a significant advantage over using only local data

model EXAM



Dayan, I., Roth, H.R., Zhong, A. et al. Federated learning for predicting clinical outcomes in patients with COVID-19. Nat Med 27, 1735–1743 (2021). https://doi-org.ubu-es.idm.oclc.org/10.1038/s41591-021-01506-3