DATA LAKE:

Un data lake es un repositorio de almacenamiento que contienen una gran cantidad de datos en bruto almacenados individualmente con un identificador único y se etiqueta con un conjunto de etiquetas de metadatos extendidas.

Esto permite obtener no sólo el dato en sí, sino también los datos que estén relacionados con el datos objeto de consulta. El principal beneficio de un data lake es la centralización de fuentes de contenido dispares. Una vez reunidas estas fuentes pueden ser combinadas y procesadas utilizando big data, búsquedas y análisis que de otro modo hubieran sido imposibles.

Recientemente el Gobierno de España anunció un Plan Estratégico para la Recuperación y Transformación Económica (PERTE) dedicada a la "Salud de Vanguardia"4. Unos de los objetivos de este plan es la construcción de una Data Lake sanitario para el Sistema Nacional de Salud , descrito como un centro de datos sanitarios que recoja la información de los sistemas de información y permita un análisis masivo para la identificación y mejora del diagnóstico y de los tratamientos. Esta infraestructura se iniciaría en el último trimestre de 2021 y concluiría en el último trimestre de 2025, con una inversión asignada de 100 millones de euros.

La empresa española **Savana**, dedicada a la transformación de los registros médicos electrónicos en información útil para su procesado m**ediante tecnologías de PLN y ML,** ha presentado una inciativa apoyada por numerosas sociedades médicas y asociaciones de pacinetes para la instalación de dicha infraestructura. El objetivo es crear un sistema interoperable de bases de datos clínicas federadas orientado a investigación sanitaria, obtenido al procesar con Machine Learning (ML) las Historias Clínicas Electrónicas de los hospitales y centros de atención primaria de los diferentes Servicios de Salud. Facilitando una investigación clínica más robusta, asequible, multicéntrica e integral.

La cuestión acerca de la privacidad de los datos se solucionaría mediante la seudoanonimización o anonimización de los mismos, como exige el Reglamento General de Protección de Datos. Según representantes de esta empresa, esta infraestructura sería la base de datos de información clínica más grande del mundo y una iniciativa pionera a nivel mundial.