Componentes de software y documentación del caso de uso de economía digital

MLEDGE - Aprendizaje automático en la nube y en el borde

(Cloud and Edge Machine Learning)

Diciembre de 2024

Información sobre el entregable

Nombre del documento:

M2 - Componentes de software y documentación del caso de uso de economía digital

Versión actual: 1.0

Proyecto: MLEDGE - Aprendizaje automático en la nube y en el borde (Cloud and Edge Machine Learning)

Paquete de trabajo: P3 - Implementación del caso de uso de economía digital

Tareas: El entregable es resultado del trabajo en los diversos componentes técnicos:
- A4.2: Implementación y pruebas del caso de uso de economía digital

Entregable: E3.2 - Componentes software, informe y documentación preliminares.

Autores: Acuratio Europe S.L. y Orange Espagne S.A.U

Revisores: Nikolaos Laoutaris (IMDEA) y Javad Dogani (IMDEA)

Historial de Versiones

Versión	Fecha	Resumen de modificaciones
Versión 1.0	31-12-2024	Versión final del documento

1. Introducción

En diciembre de 2022, el Ministerio de Asuntos Económicos y Transformación Digital del Gobierno de España adjudicó a IMDEA Networks el proyecto titulado "*MLEDGE - Aprendizaje automático en la nube y en el borde (Cloud and Edge Machine Learning)*" (REGAGE22e00052829516, en adelante el 'Proyecto' o MLEDGE). Esta iniciativa cuenta con financiación de la Unión Europea a través del Plan de Recuperación, Transformación y Resiliencia (European Union - NextGenerationEU/PRTR).

El objetivo principal del Proyecto es establecer un ecosistema robusto de servicios de aprendizaje federado (FL, por sus siglas en inglés) en el borde, que sean seguros y eficientes. Estos servicios están diseñados para facilitar el uso de datos personales y confidenciales, tanto de tipo B2B como de consumidores, en el entrenamiento de modelos de aprendizaje automático (ML), garantizando en todo momento la privacidad de los datos y de sus propietarios.
Los **objetivos generales del proyecto** se pueden resumir en los siguientes:

1. Facilitar la accesibilidad del aprendizaje federado en el borde mediante el desarrollo de una capa de software intermedio y componentes que simplifiquen la complejidad inherente al procesamiento y al intercambio de datos.
2. Abordar los desafíos técnicos asociados al aprendizaje federado en entornos de nube y borde, optimizando su implementación y rendimiento.
3. Validar la funcionalidad desarrollada a través de casos de uso representativos de problemas reales de la industria, demostrando el impacto práctico de estas tecnologías.
4. Difundir y explotar los resultados del Proyecto, involucrando a agentes externos clave y comunicando los hallazgos a un público más amplio y potencialmente interesado.

Uno de los objetivos fundamentales del proyecto consiste en diseñar, implementar y poner a disposición del público demostradores que trabajen con datos sensibles, tanto de carácter industrial como personal. Estos demostradores están orientados a alimentar modelos de aprendizaje automático aplicables a diversos sectores de la industria. En la etapa inicial del proyecto, se seleccionaron empresas clave para el desarrollo de la plataforma FLaaS (*Federated Learning as a Service*), el monitoreo de costes computacionales, y el diseño e implementación de casos de uso reales que se beneficien del aprendizaje distribuido en el borde de la nube.

La adjudicación del paquete de trabajo P4, enfocado en el diseño e implementación del caso de uso relacionado con la economía digital, recayó en la UTE ORANGE - ACURATIO EUROPE, con NIF U70737150.

El presente documento corresponde al entregable E4.2, titulado "*Componentes de software y documentación del caso de uso de economía digital*". Su propósito principal es detallar el desarrollo e implementación de los elementos definidos en el entregable E4.1, describiendo el proceso seguido y presentando la solución preliminar alcanzada.

2. Definición del problema y objetivos

En este apartado se presenta un resumen de la definición del problema y los objetivos específicos del proyecto, tal como se detallaron en el entregable E4.1. Asimismo, se describen los casos de uso relevantes, con el objetivo de proporcionar un marco de referencia que facilite la comprensión de las tareas realizadas en este entregable y su integración en el contexto general del proyecto.

2.1. Objetivos

Con el objetivo de controlar la tasa de contacto, nuestro caso de uso tratará de identificar áreas de riesgo de epidemia con datos de movilidad de operadoras telefónicas. Para ello se calcularán zonas de riesgo de contagio, asegurando la privacidad de los individuos y los datos de cada una de las entidades colaboradoras. El objetivo final es demostrar que con el apoyo de todas las operadoras de telefonía con red propia que operan en España, se podría tener una imagen completa y en tiempo real del avance de una infección.

Para una primera aproximación se ha demostrado el potencial de la solución analizando datos de dos grandes ciudades españolas, Madrid y Barcelona. Para ello se han obtenido y procesado datos de movilidad de los clientes de Orange. Estos datos se han obtenido de los Call Detail Records (CDRs) y de los eventos recogidos de las Sondas de Red (Probes) del operador. En ambos casos se recogen los eventos que son las comunicaciones que cada dispositivo realiza con las antenas más próximas y, después de un intensivo procesamiento de datos, se generan trayectorias aproximadas (ubicación y hora) de los usuarios de los dispositivos móviles.

Este análisis está dando lugar a la generación de mapas de densidad de población que, combinados con información sobre localizaciones de hospitales y otras fuentes de datos, esperamos que pueda ayudar a los equipos de emergencia o a las autoridades sanitarias a evaluar el estrés esperado de los servicios sanitarios y a los ciudadanos a decidir auto-confinarse o evitar distintas zonas. Se profundizará en este análisis y se extenderá para detectar hospitalizaciones y auto-confinamientos en el próximo hito de este proyecto.

2.2. La solución

En este punto, se detalla la aplicación de técnicas de Aprendizaje Federado (Federated Learning) a los datos de redes móviles de telefonía para la toma de decisiones en el control de epidemias.

¿Por qué Federated Learning?

Diversas soluciones de análisis de datos de movilidad y rastreo de contactos se han propuesto, pero tanto la comunidad científica como la sociedad está preocupada por la seguridad y privacidad de estas soluciones. Por ejemplo, en abril del 2020 más de 300 expertos de 25 países pidieron evitar tecnologías que permitan “una vigilancia sin precedentes de la sociedad” en la lucha contra el COVID-19, advirtiendo que la adopción de tecnologías de vigilancia "obstaculizaría catastróficamente la aceptación por parte de la sociedad en general de las aplicaciones que pueden rastrear y frenar una segunda ola de contagios tras el confinamiento”.

En España los estudios de movilidad siempre generan noticias, debido a la sensibilidad de los datos a tratar y la cesión de estos a agencias gubernamentales, como en 2019 cuando el

Instituto Nacional de Estadística solicitó un estudio sobre 50 millones líneas móviles a los tres grandes operadores, Orange, Movistar y Vodafone.

Este tipo de soluciones, para que sean desplegadas a escala y aceptadas por la sociedad, deberán tener las más altas garantías de privacidad y confidencialidad de los datos analizados. Federated Learning en combinación con Differential Privacy y K-anonimato actualmente es la solución más prometedora, por sus beneficios en cuanto a anonimización, minimización del acceso al dato y escalabilidad para analizar Terabytes de información para actualizar los modelos diariamente.

Alcance del estudio

Para demostrar la viabilidad del enfoque federado, entender sus limitaciones y construir un demostrador, Orange ha proporcionado datos reales en varios periodos de tiempo:

- Una semana de datos de movilidad al principio de la pandemia, de los clientes de dos grandes ciudades españolas, Madrid y Barcelona. Estos datos después de un intensivo trabajo de estudio, limpieza y transformación se han convertido en trayectorias a lo largo del día de cada cliente. Estas trayectorias son una serie de pares de ubicaciones y horas a lo largo del día.
- Una semana un año después del inicio de la pandemia en las mismas ciudades.
- Una semana de datos en una fecha reciente al comienzo del estudio.

Fuentes de datos que emplear y preparación de estas

En esta sección se describen los diferentes datasets que han sido utilizados en el Proyecto para generar los entregables descritos a partir de los datos de telefonía, así como los procesos de adecuación necesarios para los mismos:

CDR y sondas de red.

Se trata de los registros que los terminales móviles registran en la Red Móvil, tanto por eventos Activos como Pasivos y que llevan asociados siempre un código de tiempo (Timestamp) así como un registro de la antena móvil en la que se producen. Esto permite asociar dicho evento a un momento del tiempo y a una localización.

Estos registros se pseudo-anonimizan, se transforman de manera que recojan únicamente la información relevante para el caso de uso al que están destinados y se almacenan en un repositorio seguro gestionado por Orange.

Red móvil (antenas)

Del mismo modo, se hace necesario disponer de un inventario actualizado con los códigos asignados a cada uno de los emplazamientos de las antenas donde se están realizando los eventos recogidos. Además, este inventario recoge la información relevante como las coordenadas del emplazamiento y la información técnica necesaria (potencia de señal, etc.). Este conjunto de ficheros incluye también un fichero de *footprint*, con la huella geométrica de cobertura estimada de cada emplazamiento en base a los datos técnicos de cada uno de ellos.

Clientes

Por último, se disponibiliza en el mismo repositorio seguro la información relativa a clientes necesaria para este estudio. Esta información únicamente incluye el identificador pseudoanonimizado necesario para el cruce con los eventos (CDR y/o Sondas), eliminándose cualquier información personal de los clientes de Orange. El identificador que se utiliza es el

IMSI de cliente pseudoanonimizado, siendo éste el único identificador personal que se incluye en los dataset.

Cálculo de Trayectorias

De la combinación de los datos anteriormente descritos, una vez seleccionados, normalizados, limpios de ruido (por ejemplo, por los efectos "rebote" dentro de la señalización en la red móvil), se procederá al cálculo de las trayectorias individuales de cada cliente para cada uno de los días dentro del período indicado para el estudio y a su puesta a disposición en un repositorio para su utilización.

Infraestructura IT

Los datos pseudoanonimizados de Orange se encuentran en una infraestructura (Cloud pública AWS) segura administrada por el propio operador. Cuenta con las herramientas necesarias a nivel de monitorización y seguridad, y es auditada tanto a nivel de seguridad como de privacidad para garantizar que se cumplen los más altos estándares de calidad en estos términos.

En esta plataforma cloud, Acuratio ha desplegado su plataforma federada, que tiene acceso a un repositorio de almacenamiento donde Orange almacena los datos objeto del estudio. Desde la plataforma se podrán diseñar analíticas y entrenar modelos.

2.3. Caso de uso

El caso de uso planteado permitirá a las administraciones públicas acceder a información actualizada, completamente agregada y anónima, sobre la movilidad en un área determinada para poder tomar decisiones de una forma más eficiente y precisa.

El objetivo principal es evitar medidas drásticas como el cierre completo de una ciudad o provincia permitiendo soluciones más quirúrgicas adaptadas a la situación epidemiológica de cada área en concreto. De este modo se pretende dotar a la administración pública de herramientas que permitan limitar el impacto económico de dichas medidas y ayudar a mitigar las consecuencias de situaciones como la pandemia del Covid-19 en el futuro.

Caso de uso	Usuario	Objetivo	Beneficio, resultado, razón del caso de uso
ANÁLISIS DE MOVILIDAD EN EMERGENCIAS	Administración pública	Mejorar la toma de decisiones en la restricción de la movilidad durante una emergencia	Acceso unificado a datos de movilidad proveniente de distintas fuentes

El escenario final contempla que uno o más operadores de telefonía puedan poner sus datos a disposición de un consumidor (la administración pública) a través de la plataforma federada desarrollada para MLEDGE.

El resultado de este caso de uso serán mapas de calor agregados que permitan evaluar el riesgo de distintas áreas geográficas en función de la movilidad detectada por los operadores de telefonía.

3. Extracción y limpieza de datos

Tal y como se expuso en la entrega 4.1, titulada **"***Estado del arte y diseño de los componentes del caso de uso de economía digital***"**, el uso de datos agregados y anónimos sobre la población, obtenidos a partir de la geolocalización de dispositivos móviles, constituye una práctica consolidada y ampliamente utilizada en diversos sectores, tanto públicos como privados. Esta metodología se basa en el procesamiento responsable de la información anonimizada, lo que permite generar indicadores estadísticos derivados del análisis del comportamiento de los usuarios y su localización dentro de la red móvil del operador. Desde el año 2016, Orange ha sido proveedor de este tipo de soluciones en España, aplicándolas en diferentes áreas estratégicas.

Las soluciones Smart Data de Orange ofrecen un conjunto avanzado de herramientas analíticas diseñadas para extraer conocimiento valioso sobre los patrones de movilidad y el comportamiento generalizado de la población. Estas herramientas permiten transformar los registros de señalización de usuarios móviles, siempre en un formato anónimo y cumpliendo estrictamente con la normativa de protección de datos, en indicadores estadísticos clave. Ejemplos de estos indicadores incluyen la frecuencia de visitas a determinadas zonas geográficas, la intensidad de los desplazamientos entre áreas específicas y las tendencias de movilidad en diferentes momentos del tiempo.

Para la realización del presente estudio, Orange está proporcionando datos de movilidad detallados correspondientes a las ciudades de Madrid y Barcelona. Estos datos abarcan tres periodos específicos: una semana al inicio de la pandemia, una semana transcurrido un año desde el inicio de esta y, finalmente, una semana correspondiente a una fecha más reciente, seleccionada al comienzo del estudio.

A continuación, se detalla el volumen final de los datos extraídos, después de la limpieza y procesamiento de los mismos.

Periodo	Año	Volumen de datos Final
Primera semana de marzo	2020	40,6 Gb
Primera semana de marzo	2021	38 Gb
Primera semana de marzo	2024	53,2 Gb

3.1. Extracción de datos

Costes del proceso

Para la extracción de los datos de trayectorias se han utilizado cuatro fuentes de datos pertenecientes a Orange: geolocalizaciones, datos de red, datos de clientes, y cobertura de las antenas de telefonía. Todos estos datos suman un volumen inmenso, en el orden de los cientos de GB diarios. Para poder procesarlos y extraer las trayectorias de interés ha sido necesario movilizar recursos de Orange tanto técnicos como de infraestructura en la nube.

Uno de los principales desafíos ha sido el volumen de los datos, para solventarlo se han utilizado herramientas especializadas de procesamiento en la nube que se describen con detalle en el punto 3.2, tales como: AWS Athena, Scala o Spark.

También ha sido necesario prestar especial atención a los procesos de anonimización al tratarse de datos sensibles como la ubicación de los clientes de Orange. Para lograr una anonimización efectiva se ha aplicado un hasheo sobre el identificador con una clave controlada por Orange, este es un proceso habitual en su operativa y de hecho los datos suelen almacenarse pseudo-anonimizados en origen. Adicionalmente, tal y como se explicará a continuación, el uso de geohashes con una precisión limitada facilita la anonimización al no ser posible localizar un identificador de forma exacta,

En la siguiente tabla se puede ver un desglose de los procesos y el tiempo que se ha invertido en ellos.

#	Acciones	Tiempo
Desarrollo	• Aparición y configuración del software	10 días
Copiado input	• Análisis • Creación de infraestructuras • Permiso para copiado SoiProd-> Squad1 • Copiado	10 días
Ejecución del Proceso	• Activación del Entorno • Despliegue Software • Realizar ejecución • Validación de datos	32 horas Aprox. (1h 30mins por día de ejecución)
Copiado output	• Permisos para copiado Squad1 -> Partner3 • Copiado	4 días

3.2. Preparación de datos

Los datos utilizados incluyen localizaciones diversas, como las correspondientes a CGI, hospitales y domicilios de usuarios, representadas en distintos formatos. Para unificar y estandarizar la representación de estas localizaciones, se emplea la codificación mediante Geohash.

3.2.1. Codificación Geohash

Geohash es un sistema de codificación espacial de dominio público diseñado para representar ubicaciones geográficas de manera compacta mediante cadenas de texto alfanuméricas. Este sistema permite dividir la superficie terrestre en una serie de celdas de tamaño variable, proporcionando una forma eficiente de gestionar datos geoespaciales.

El Geohash convierte las coordenadas geográficas de latitud y longitud en cadenas alfanuméricas, logrando esta conversión al subdividir el mapa en 32 celdas, cada una representada por un carácter: los números del 0 al 9 y las letras b, c, d, e, f, g, h, j, k, m, n, p, q, r, s, t, u, v, w, x, y, z. Estas celdas pueden subdividirse recursivamente en otras 32 celdas más pequeñas, permitiendo alcanzar el nivel de precisión requerido.

La longitud de la cadena Geohash determina su precisión: a mayor longitud, mayor nivel de detalle, ya que delimita un área geográfica más reducida.

Por ejemplo:

- 5 caracteres pueden representar un área de 12.5km^2 aproximadamente.
- 6 caracteres pueden representar un área de 0.7km^2 aproximadamente.

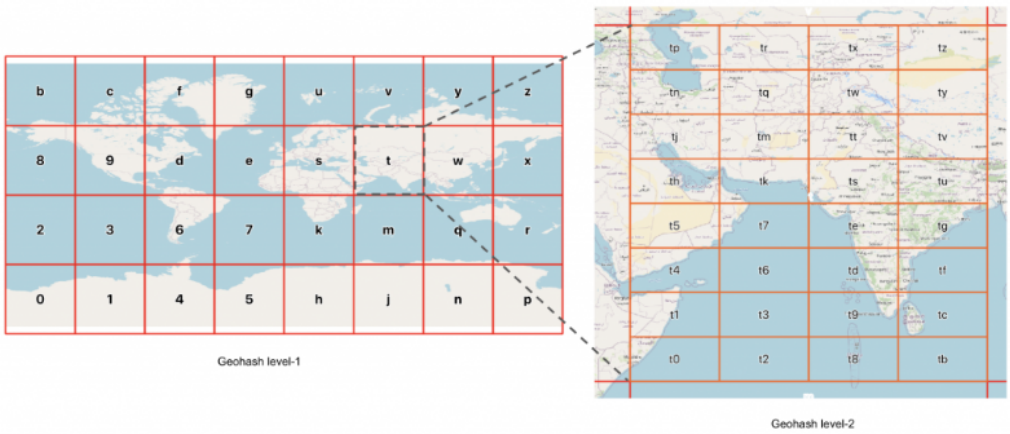

Ilustración 1: Ejemplo geohash

El área que abarca un geohash con un número de caracteres concreto puede variar según la latitud, ya que las celdas de geohash son más grandes cerca del ecuador y se hacen más pequeñas hacia los polos debido a la convergencia de las líneas de longitud. En este estudio, al tratarse de geolocalizaciones en una misma ciudad, la variación entre distintos geohash con el mismo número de caracteres es despreciable.

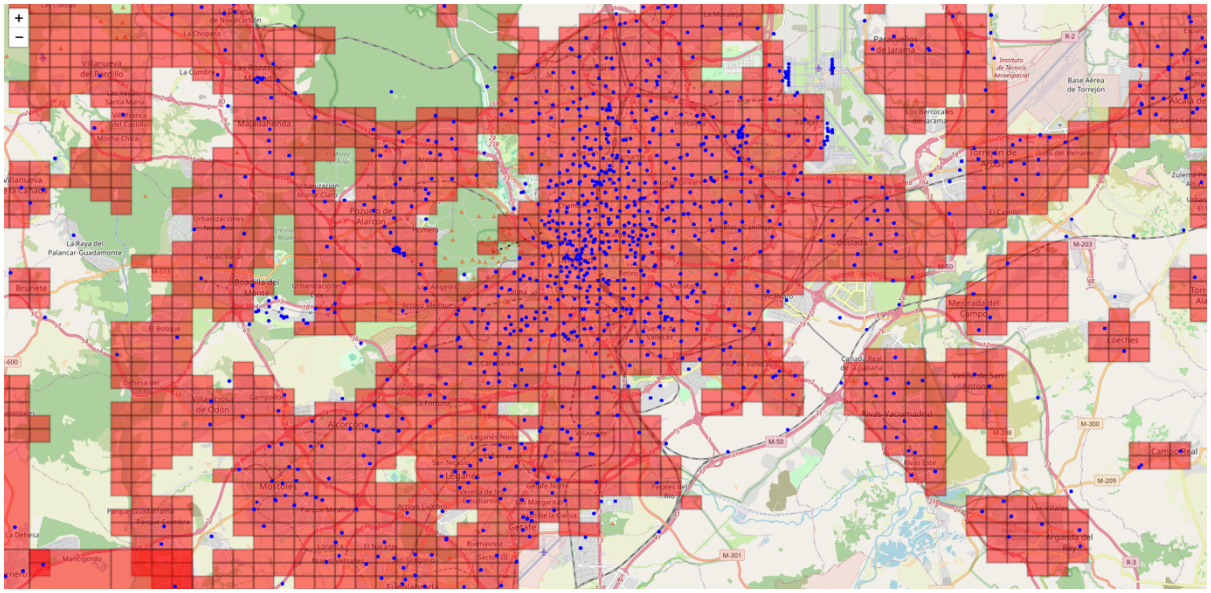

Ilustración 2: Ejemplo de Geohash aplicados en este proyecto y distribución de las ubicaciones de las antenas (Red Móvil). Madrid.

3.2.2. Tratamiento de datos

Los datos iniciales en crudo están almacenados en el Data Lake de Orange en AWS y el primer paso del proceso es la definición de los periodos iniciales de interés y el copiado de estos datos a la infraestructura de procesamiento. Una vez copiados estos datos a la infraestructura de procesamiento hay que transformarlos, ya que en origen se trata de grandes volúmenes de registros en formato CSV y comprimidos a formatos gzip y bzip2.

La exploración inicial de los datos se hizo en Amazon Athena, un servicio de consulta interactiva completamente administrado por AWS que facilita el análisis de grandes volúmenes de datos almacenados en Amazon S3 (Simple Storage Service) utilizando SQL estándar. Athena permite realizar consultas en datos no estructurados, como archivos CSV, JSON, Parquet, ORC, y otros formatos, sin necesidad de realizar ningún tipo de infraestructura o configuración previa.

Amazon Athena es una herramienta poderosa, fácil de usar y económica para realizar consultas SQL directamente sobre datos almacenados en Amazon S3. Es ideal para análisis interactivos, exploración de grandes volúmenes de datos y casos de uso ad-hoc, sin la necesidad de gestionar infraestructura o mover los datos a otro sistema. Además, su modelo de precios basado en el volumen de datos escaneados lo hace adecuado para proyectos de análisis de datos a gran escala.

El funcionamiento de Athena consta de 3 pasos:

1. **Definición de tablas:** En Athena, defines tablas que apuntan a los datos almacenados en S3. Esto se hace especificando la ubicación de los archivos en S3 y el formato de los datos (CSV, Parquet, etc.). Puedes usar AWS Glue para facilitar la creación de estas tablas.
2. **Consulta SQL:** Una vez que tienes las tablas definidas, puedes realizar consultas SQL estándar sobre esos datos. Athena usa Presto, un motor de consultas SQL distribuido, para ejecutar las consultas de manera rápida y eficiente.

3. **Resultados:** Los resultados de las consultas se pueden ver directamente en la consola de Athena o exportarse a otro servicio de AWS como Amazon S3, AWS QuickSight (para visualización de datos) o Amazon Redshift (para almacenamiento de datos).

Ilustración 3: Arquitectura Amazon Athena.

De esta exploración inicial se concluye que se pueden usar los conjuntos de datos de geolocalización y red para construir las trayectorias de los clientes. Al conocer el CGI (Cell Global Identity) al que está conectado cada usuario, es posible determinar con precisión el área geográfica en la que se encuentra en un momento dado. Esto no solo permite mapear la ubicación de cada usuario, sino también calcular el tiempo que permanece en cada área, facilitando así el análisis de sus trayectorias y comportamientos de movilidad.

A partir de estos datos en crudo se generan dos conjuntos de datos de entrada que servirán para identificar estas trayectorias:

1. Camino recorrido por un dispositivo móvil: Contiene los datos de las posiciones geográficas (geohashes) y los tiempos en que se registraron. Formato:
    - *telco_id*: Identifica al proveedor o usuario (pseudo-anonimizado).
    - *geohash* (A) y *geohash* (B): Representan una transición de un punto geográfico a otro.
    - *id* (session): Sesión única del usuario.
    - *timestamps_info*: Tiempos asociados al movimiento.
2. Transiciones de geohashes: Representan los cambios entre un geohash origen (*geohash A*) y un geohash destino (*geohash B*).

La tabla de *tracks* se genera cruzando ambos conjuntos de datos. Esto permite identificar patrones de movimiento y calcular métricas adicionales (como tiempo entre transiciones, velocidad, etc.).

Todo este proceso se realiza utilizando Scala, un lenguaje de programación de propósito general diseñado para ser conciso, expresivo y flexible, combinando características de la programación funcional y la programación orientada a objetos.

Explicación del código

Función principal:

La función *getFinalTracksDf* genera un *DataFrame* final que incluye información enriquecida sobre los movimientos de un usuario a partir de dos entradas:

- *eventsDF*: Detalles de eventos del usuario (posiciones y tiempos).
- *tracksDF*: Información de transiciones de geohashes.

Creación de identificador de track:

> *.withColumn(conf.trackIdField, concat(col(conf.originalGeohashField), lit(" "), col(conf.nextGeohashField)))*

Se genera un campo único (*track ID*) concatenando el geohash A y el geohash B. Esto identifica una transición específica entre dos puntos.

Unión con la tabla de transiciones:

> *.join(tracksDF, conf.trackIdField)*

Se cruza la información de eventos (*eventsDF*) con la tabla de transiciones (*tracksDF*) utilizando el *track ID*. Esto combina detalles de ambas fuentes.

Cálculo de métricas por sesión:

Se define una ventana de partición por usuario y sesión:

> *val telcoSessionWindow = Window.partitionBy(conf.telcoField, conf.sessionIdField)*

Luego, se agregan columnas calculadas:

1. Tiempo transcurrido (*elapsedTimeField*): El tiempo entre dos eventos consecutivos se calcula como:

   toLong(conf.nextTimestampField) - toLong(conf.lastTimestampField)

   Si es el primer geohash de la sesión, usa el tiempo desde el inicio de la sesión:

   toLong(conf.lastTimestampField) - toLong(conf.startTimestampField)

2. Identificación de movimiento motorizado (*motorField*): Se marca como true si la velocidad registrada (*tripSpeedField*) supera los 50 km/h:

   .withColumn(conf.motorField, when(col(conf.tripSpeedField) > 50, true).otherwise(false))

3. Información de tiempo:
   - Mes: Calculado a partir de la fecha de inicio.
   - Semana del año: Identifica la semana del movimiento.
   - Tipo de día (*dayField*): Determina si el movimiento ocurrió en fin de semana (“*weekend*”) o día laborable (“*business_day*”).
   - Horario del día (*scheduleField*): Clasifica el evento en:
     - *“morning”* (6:00 a 13:00).
     - *“afternoon”* (14:00 a 21:00).
     - *“night”* (otros horarios).

Las trayectorias generadas mediante este proceso deben limpiarse para tener en cuenta anomalías tales como clientes que se encuentran en varias ubicaciones diferentes al mismo

tiempo, o clientes que viajan cientos de kilómetros en cuestión de minutos. Estas anomalías ocurren porque los eventos de red son extremadamente frecuentes y puede haber pequeños errores en los datos recogidos, es necesario identificar y eliminar estos casos extraños ya que podrían influir negativamente en las conclusiones definitivas.

Una vez obtenidas las trayectorias de los usuarios, los datos se cruzan con el archivo *clientes*, que contiene información esencial sobre los clientes, tales como su ID único, la ubicación de su hogar y su lugar de trabajo. Esta información es fundamental para realizar un análisis profundo de los patrones de movilidad de los usuarios.

El archivo *clientes* contiene, entre otros, los siguientes datos:

- *telco_id*: Un identificador único que permite reconocer al cliente dentro del sistema de telecomunicaciones.
- *postal_code_id*: El código postal de la dirección donde reside el cliente.
- *age_des*: La edad del cliente.
- gender_des: Género del cliente.
- *customer_type_des*: El tipo de cliente, que puede variar según las características del servicio contratado o el perfil del usuario.
- *locality_des*: La localidad en la que reside el cliente.
- *operator*: El operador con el que el cliente tiene contratado el servicio.
- *cgi_home*: El CGI (Cell Global Identity) de la antena de telecomunicaciones más cercana a su hogar.
- *cgi_work:* El CGI de la antena de telecomunicaciones más cercana a su lugar de trabajo.

De esta tabla los datos de interés son las localizaciones de trabajo y hogar de cada cliente. Para realizar el proceso de forma más eficiente se decide cruzar primero la tabla de *clientes* con las tablas de trayectorias en las fechas seleccionadas, de este modo trabajaremos solo sobre el subconjunto de clientes que nos interesan.

Una vez seleccionados estos clientes, los datos se cruzan con el archivo *footprint*, el cual contiene las huellas geográficas de las antenas de telecomunicaciones, utilizando el formato GeoJSON. Cada registro en este archivo asocia un CGI con su correspondiente área de cobertura y geometría, lo que permite mapear visualmente las áreas cubiertas por cada antena. Esta información es crucial para entender el contexto espacial de los datos de geolocalización, ya que permite determinar en qué área se encuentra el usuario en cada momento, lo que resulta fundamental para un análisis detallado de sus patrones de movilidad.

El archivo contiene los siguientes datos:

- *cgi*: El identificador único de la antena de telecomunicaciones.
- *maxx*: El límite superior del área de cobertura en el eje X.
- *minx*: El límite inferior del área de cobertura en el eje X.
- *maxy*: El límite superior del área de cobertura en el eje Y.
- *miny*: El límite inferior del área de cobertura en el eje Y.
- *geometry*: La representación geométrica del área de cobertura, en la que se especifican los límites máximos y mínimos de las coordenadas.

El resultado de este cruce permite identificar los geohash del hogar y el lugar de trabajo de los clientes junto con sus trayectorias diarias, datos que serán esenciales para identificar

hospitalizaciones. Este proceso de cruce se lleva a cabo en Amazon Athena, ya que el volumen de datos se ha reducido significativamente en este punto. La siguiente consulta SQL ilustra cómo se realiza este proceso:

```
SELECT DISTINCT "default"."tracks20240224"."telco_id",
cgi,
minx,
maxx,
miny,
maxy
FROM "default"."clientes"
INNER JOIN "default"."footprint" ON "default"."clientes"."cgi_home" =
"default"."footprint"."cgi"
INNER JOIN "default"."tracks20240224" ON "default"."clientes"."telco_id" =
"default"."tracks20240224"."telco_id"
WHERE "default"."clientes"."cgi_home" IS NOT NULL;
```

En él se realizan los siguientes pasos.

Primero se seleccionan las siguientes columnas; *teleco_id, cgi, minx, maxx, miny, maxy,* que se corresponden a las variables de *footprint* antes explicadas. Estableciendo que no contengan filas duplicadas, lo que significa que solo se devolverán combinaciones únicas de las columnas seleccionadas.

La consulta trabaja con tres tablas. *Clientes* (que contiene datos como *telco_id* y *cgi_home*), *footprint* (que contiene las variables antes mencionadas) y *tracks20240224* (que contiene información de los movimientos de los usuarios, incluyendo el *telco_id* de cada cliente y los datos relacionados con sus movimientos en la red del día 24 de febrero de 2024).

Es importante señalar que, para determinar el CGI de la vivienda y el del trabajo, se ha utilizado un historial más extenso, siguiendo el criterio que se detalla en la imagen a continuación. Para calcular el “cgi_home”, se considera el CGI al que el usuario ha permanecido conectado durante la noche. Por otro lado, para establecer el “cgi_work”, se toma en cuenta el CGI al que el usuario se conecta entre las 10:00 a.m. y las 2:00 p.m. durante los días laborables.

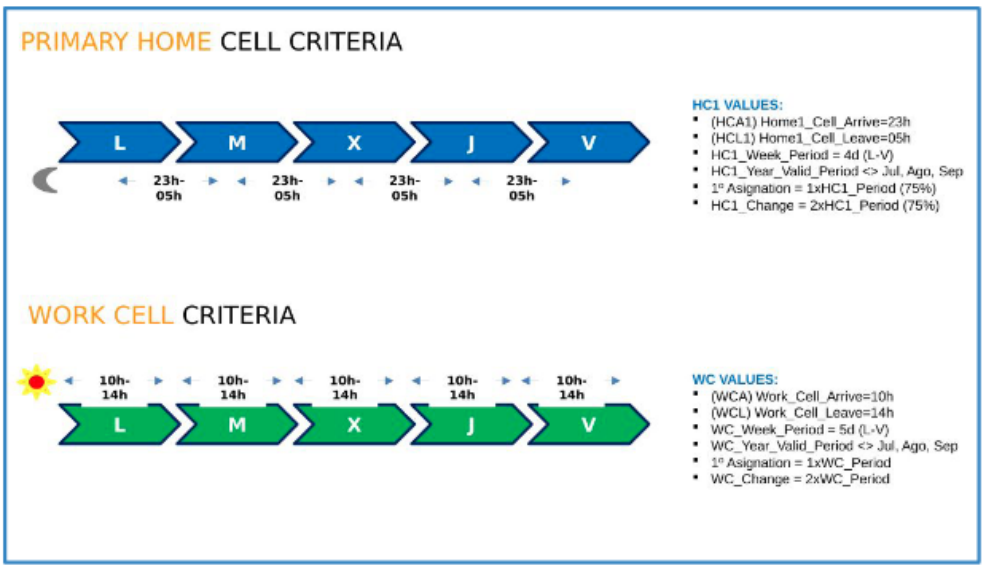

Ilustración 4: Criterio selección CGI vivienda y trabajo.

Se han unido las tablas *clientes* y *footprint* en los casos donde el *cgi_home* de los clientes coincide con el CGI en la tabla *footprint*. Esto asocia a cada cliente con la antena de telecomunicaciones correspondiente a su hogar. Luego se unen con la tabla *trascks20240224* usando el *telco_id*, que debe coincidir en ambas tablas, lo que asocia los movimientos de cada cliente con la información del cliente.

Por último, se establece la condición de la consulta, filtrando solo los clientes cuyo *cgi_home* no sea nulo. Esto implica que sólo se seleccionan los clientes que tienen asignada una antena de telecomunicaciones en su hogar.

En resumen, la consulta selecciona los *telco_id* de los clientes, junto con la huella geográfica (*minx, maxx, miny, maxy*) de las antenas asociadas a sus hogares, filtrando para que solo se consideren los clientes que tienen un *cgi_home* (antena en su hogar) definido. Mediante un proceso similar se realiza la consulta para *cgi_work*, estos se realizan para las trayectorias de cada día del que disponemos datos.

Como se ha mencionado previamente, con el fin de reducir y compactar los datos de localización, estos se codifican utilizando Geohash. Esta técnica asigna una referencia única de 6 caracteres a cada localización, lo que permite simplificar y optimizar el análisis espacial.

El proceso de codificación a Geohash se lleva a cabo en *python,* utilizando la librería *pygeohash*, que ofrece una implementación sencilla para codificar y decodificar geohashes. Al introducir una tupla de latitud y longitud junto con el nivel de precisión deseado (en este caso, 6), se obtiene un geohash que representa de manera eficiente la ubicación del usuario.

En resumen, a partir de los datos obtenidos de la combinación de las trayectorias de los usuarios, la información de los clientes y las huellas geográficas de las antenas se puede determinar con gran precisión qué usuario se encuentra en qué área, y durante cuánto tiempo, permitiendo realizar un análisis exhaustivo de los patrones de movilidad.

3.2.3. Requisitos

Finalmente, recogemos a continuación los requisitos que debían cumplir los módulos de limpieza y extracción de datos, y aportamos la justificación correspondiente:

RF1	Exploración	El sistema debe permitir explorar los datos en crudo con un coste controlado, por ejemplo, utilizando soluciones *serverless*	Obligatorio
RF2	Preprocesamiento de datos crudos	El sistema debe permitir procesar los datos en crudo para obtener las trayectorias necesarias para el desarrollo del caso de uso.	Obligatorio

Tal y como se ha descrito, el procesamiento se ha hecho utilizando Spark y Amazon Athena, lo que ha permitido extraer la información necesaria para los casos de uso planteados minimizando los costes. Tal y como se dijo en el anterior entregable se han utilizado soluciones *serverless* para procesar los datos en crudo y obtener las trayectorias necesarias.

RF3	Estandarización	El sistema debe definir un formato común para las tablas procesadas con las trayectorias de los clientes que permita la agregación de estos resultados mediante la plataforma federada	Obligatorio

Se ha descrito un formato único para las trayectorias, y al haber obtenido este formato a partir de los datos de red en crudo proporcionados por Orange el proceso debería ser replicable con los datos de otras operadoras.

RF4	Almacenamiento	El sistema debe almacenar las trayectorias generadas para que la plataforma federada pueda acceder a ellas. Este almacenamiento debería ser cifrado en reposo para garantizar la seguridad	Obligatorio
RF5	Análisis	El sistema debe permitir analizar las trayectorias para generar las capas de visualización descritas en el caso de uso.	Obligatorio

El almacenamiento de las trayectorias generadas en S3 cumple estos dos criterios ya que garantiza el cifrado en reposo de los datos y además los disponibiliza para que la capa de ejecución pueda acceder a ellos. Además, al tener estos datos accesibles desde los nodos de la plataforma se podrán analizar las trayectorias para generar las capas de visualización descritas.

3.3. Creación de Bases de Datos

3.3.1. Base de datos de Hospitales

En primer lugar, se recopiló un listado de hospitales públicos y privados ubicados en Madrid y Barcelona utilizando una base de datos pública[1] proporcionada por el Ministerio de Sanidad. Esta base incluye información detallada sobre los hospitales de cada comunidad autónoma, como el municipio, la dirección y otros datos relevantes.

A continuación, se verificó la participación de estos hospitales en la hospitalización de pacientes infectados durante la pandemia de COVID-19. Tras este análisis, se excluyeron aquellos hospitales que no atendieron a pacientes infectados, conservando únicamente los que sí estuvieron involucrados.

Posteriormente, se determinó la ubicación geográfica de los hospitales seleccionados, codificando sus coordenadas mediante el sistema Geohash. Esta codificación permite facilitar la comparación entre el *geohash_id* de un usuario y el *geohash_id* de un hospital en particular.

[1] https://www.sanidad.gob.es/ciudadanos/centrosCA.do

Finalmente, se elaboró una muestra de la base de datos correspondiente a los hospitales de Barcelona. Esta muestra, generada tras aplicar los filtros mencionados, incluye información adicional relevante, como las coordenadas en formato Geohash, para facilitar un análisis más detallado.

	A	B	C	D
1	Hospitales	(Latitud,Longitud)	Geohash 6	Aceptaban Pacientes COVID-18
2	Hospital Municipal de Badalona	(41.45048532985253, 2.245606997136341)	sp3eg1	YES
3	Hestia Palau.	(41.41010057828765, 2.1724940259704244)	sp3e96	YES
4	Hospital Universitari Sagrat Cor	(41.388849024599736, 2.145737364417767)	sp3e2y	YES
5	Clinica Mc Londres	(41.38943446880813, 2.1455200587787795)	sp3e2y	YES
6	Clínica Tres Torres (cirugia privada)			NO (cirugías privadas)
7	Hospital de La Santa Creu I Sant Pau	(41.41549442552108, 2.174704239464017)	sp3e97	YES
8	Hospital Quirónsalud Barcelona	(41.415669079666216, 2.1385764259706264)	sp3e8e	YES
9	Hm Nou Delfos	(41.41348634064367, 2.1424870327172445)	sp3e8g	YES
10	Hospital El Pilar	(41.40036822266783, 2.1484202775426886)	sp3e8b	YES
11	Centre Hospitalari Policlinica Barcelona	(41.40327721366414, 2.149555312476774)	sp3e8c	YES
12	Serveis Clinics, S.A.	(41.41249973167111, 2.1360053415434854)	sp3e8d	YES
13	Clínica Coroleu - Ssr Hestia.			NO (Hospital de salud mental)

Ilustración 5: Muestra ubicaciones hospitales Barcelona.

Además, se presenta un mapa de Madrid y Barcelona que destaca los geohash correspondientes a la ubicación de los hospitales. Esta representación visual permite identificar y analizar de manera sencilla las áreas con mayor densidad hospitalaria.

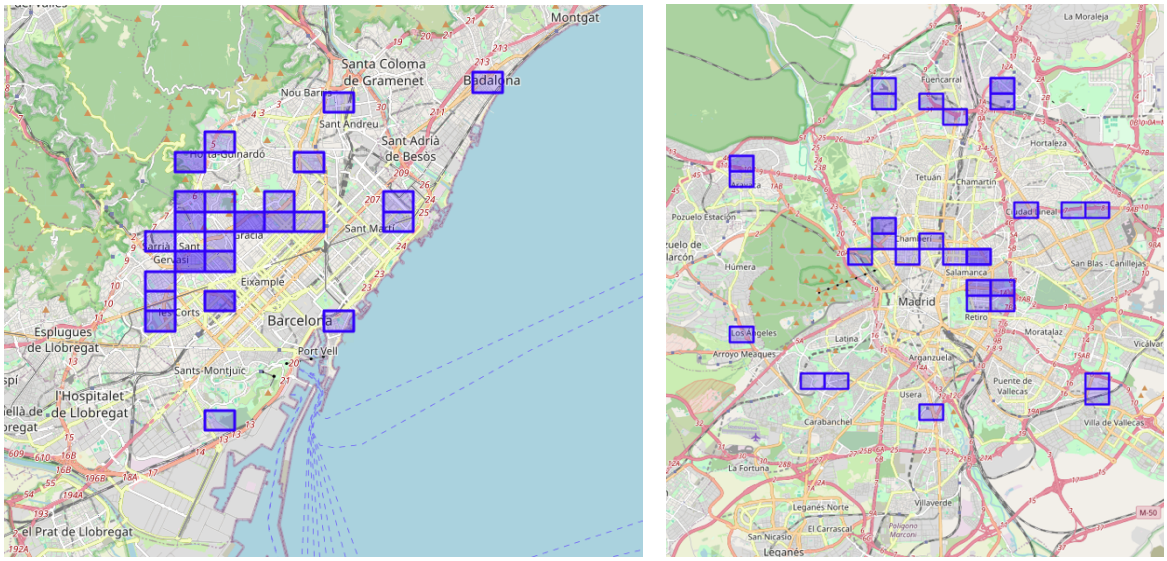

Ilustración 6: Geohases Hospitales Madrid y Barcelona

3.3.2. Base de datos de Madrid y Barcelona

Una vez completado el paso anterior, se procede a filtrar la información con el objetivo de generar una base de datos optimizada para el análisis. Este proceso comienza con una selección inicial de los datos, agrupando los identificadores de Geohash (*geohash_id*) según sus tres primeros caracteres.

En esta etapa preliminar, se excluyen los Geohash que no comienzan con el prefijo "ezj" en la base de datos de Madrid, ya que este prefijo corresponde a las áreas geográficas que abarcan la región de Madrid. De manera similar, en la base de datos de Barcelona, se descartan los Geohash que no inician con "sp3", prefijo que delimita las zonas correspondientes a dicha ciudad.

Posteriormente, se realiza un segundo filtrado más específico. Para la región de Madrid, se seleccionan únicamente los Geohash con prefijos "ezjq" y "ezjm", que abarcan la mayor parte del territorio. En el caso de Barcelona, se incluyen los Geohash con prefijos "sp36", "sp37" y "sp3e", los cuales cubren las principales áreas urbanas de la ciudad.

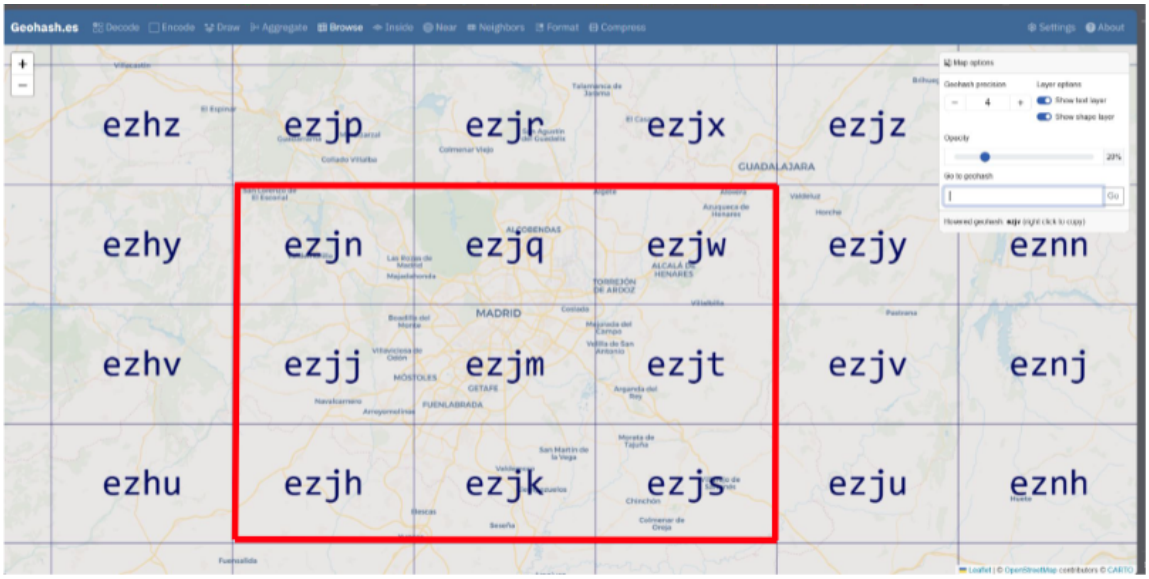

Ilustración 7: Geohashes seleccionados para Madrid.

De cara al próximo entregable, se considera la posibilidad de reducir el área de trabajo para ajustarla de forma más precisa a la ciudad de Barcelona, optimizando así el análisis.

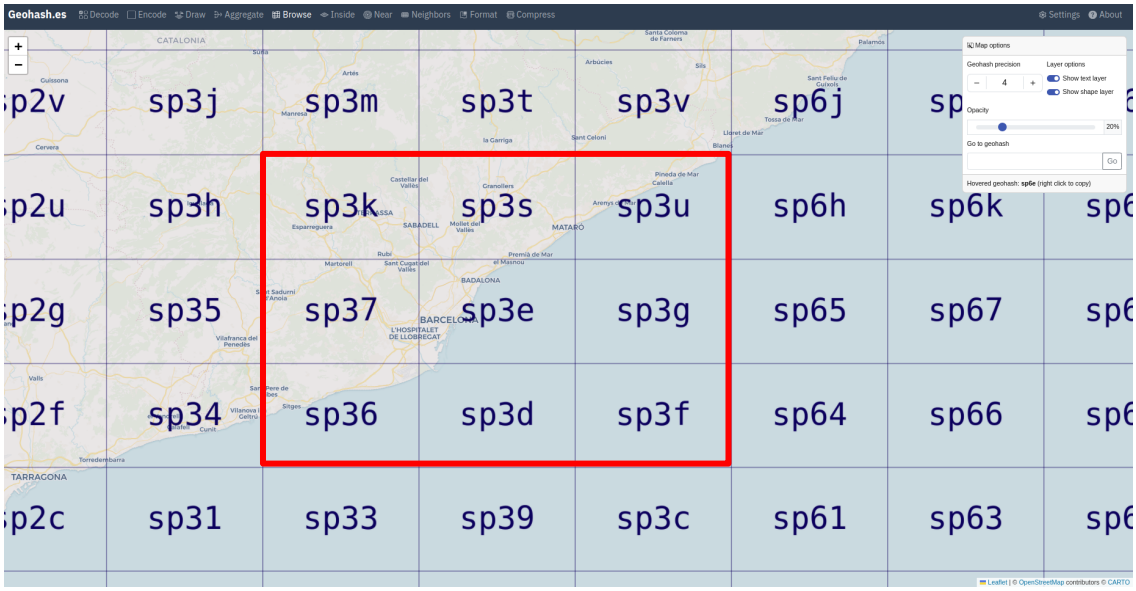

Ilustración 8: Geohashes seleccionados para Barcelona.

3.3.2. Identificación de hospitalizaciones

Como se detalla en la sección 2.1 (Objetivos), uno de los objetivos principales del proyecto es monitorear la tasa de contacto, centrándose en la identificación de áreas de riesgo

epidémico a través del análisis de datos de movilidad proporcionados por las operadoras telefónicas. Este enfoque tiene como propósito calcular zonas de riesgo de contagio, asegurando en todo momento la privacidad de los individuos y de los datos de las entidades colaboradoras.

Para realizar este análisis, una vez calculadas las trayectorias de los usuarios, estas se comparan con las ubicaciones de los hospitales para estimar el flujo de personas que los visitan, así como el número de hospitalizados en cada hospital. Esta información es fundamental para identificar áreas de alto riesgo y diseñar medidas preventivas que reduzcan la probabilidad de colapsos hospitalarios.

Tras este mapeo y la verificación inicial, los datos se someten a un nuevo proceso de filtrado. En esta etapa, se excluyen aquellos usuarios cuya ubicación residencial o laboral coincide con el *geohash_id* de un hospital, así como aquellos para los que no se dispone de información sobre su lugar de residencia o trabajo. Este procedimiento permite obtener una base de datos depurada, compuesta exclusivamente por usuarios que no residen ni trabajan en un Geohash asociado a un hospital y cuya ubicación es conocida con precisión.

La base de datos resultante representa un conjunto optimizado de usuarios cuya información geográfica ha sido cuidadosamente procesada. Este conjunto permitirá realizar análisis más específicos sobre las áreas de riesgo de contagio y los flujos de personas en relación con los hospitales, contribuyendo a una gestión más eficiente del riesgo en contextos epidémicos.

La base de datos con la que se trabajará a partir de este momento se compone de los siguientes campos:

- *telco_id*: Identificador único del usuario.
- *geohash_id*: Identificador geohash correspondiente a la ubicación del usuario en un momento determinado.
- *month*: Número entero que representa el mes de los datos registrados.
- *week*: Número entero que indica la semana específica del año a la que corresponden los datos.
- *day*: Número entero que señala el día del mes de los datos.
- *schedule*: Categoría que indica el intervalo del día al que corresponde la actividad del usuario, ya sea mañana, tarde o noche.
- *elapsed_time*: Número real que refleja el tiempo que el usuario ha permanecido en el geohash durante el periodo registrado.
- *motor*: Campo booleano que indica si el usuario estaba utilizando algún tipo de transporte motorizado. Es “True” si el usuario va en un vehículo motorizado y “False” en caso contrario.

Con estos datos, es posible determinar el tiempo exacto que un usuario ha permanecido en un geohash específico en una fecha y hora determinadas, así como conocer si el usuario estaba en movimiento utilizando un vehículo motorizado. Este tipo de información es clave para el análisis de patrones de movilidad y la identificación de zonas de riesgo, permitiendo una mejor comprensión del comportamiento de los usuarios en relación con su entorno y con las ubicaciones de interés, como los hospitales, durante el periodo de estudio.

A continuación, se presenta un ejemplo de la estructura de la base de datos. Para este ejemplo se han sustituido los identificadores para preservar la anonimidad de los usuarios, al igual que el intervalo de tiempo.

	teleco_id	geohash_id	month	week	day	schedule	elapsed_time	motor
0	aHgwldDB/oFz	ezj1s0	2	8	weekend	night	13.595714	False
1	aHw4ltDF+oFy	ezj1s1	2	8	weekend	night	45.153903	False
2	aHkztFDo/4N1	ezj1s3	2	8	weekend	night	32.114728	False
3	aHlzytDK+5ny	ezj1s5	2	8	weekend	night	5.762409	True

Ilustración 9: Muestra base de datos trayectorias.

Para realizar este análisis se han tomado referencia los hallazgos del artículo "*Using mobile network data to color risk maps*[2]", que aborda el reto de mejorar la predicción y el monitoreo de epidemias, particularmente en contextos donde los sistemas de salud pública tradicionales enfrentan dificultades para recopilar y analizar datos en tiempo real. El principal desafío radica en la necesidad de identificar con rapidez y precisión las áreas con mayor riesgo de propagación de enfermedades, un aspecto crucial para implementar medidas preventivas eficaces. Sin embargo, los métodos convencionales de vigilancia, como los informes de casos o las encuestas de movilidad, suelen ser lentos y limitados en su alcance.

Como alternativa, los autores del artículo proponen un enfoque basado en el uso de datos anónimos y agregados obtenidos de las redes móviles para mapear los patrones de movilidad de la población. Estos datos, generados a partir de las interacciones de los dispositivos móviles con las torres de señal, ofrecen una visión detallada de los desplazamientos de las personas entre diferentes áreas geográficas, proporcionando una herramienta poderosa para identificar y mitigar riesgos en tiempo real.

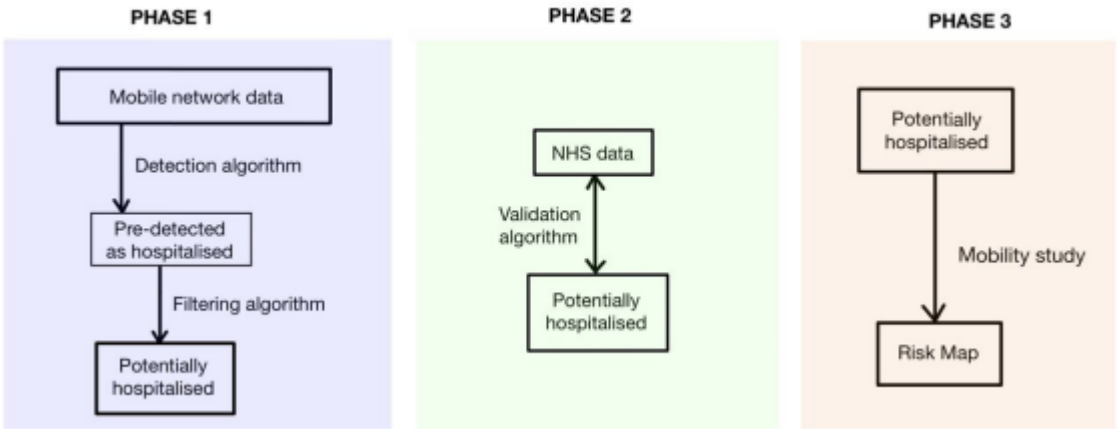

Ilustración 10: Método indentificación hospitalizaciones.

A través del análisis de estos flujos de personas, los investigadores pueden identificar con mayor precisión las áreas con mayor riesgo de contagio, sin necesidad de depender de métodos tradicionales que suelen ser más lentos o limitados en su alcance. Al contar con una fuente de datos que proporciona información casi en tiempo real, este enfoque permite a las autoridades sanitarias monitorear de manera continua y dinámica la movilidad de la población,

[2] E. Cabana, A. Lutu, E. Frias-Martinez, and N. Laoutaris. Using mobile network data to color epidemic risk maps. In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Spatial Computing for Epidemiology, pages 35--44, 2022

lo que a su vez facilita la detección temprana de posibles puntos de transmisión. Los mapas generados con estos datos ofrecen un panorama más detallado y actualizado, lo que mejora la capacidad de tomar decisiones informadas y de implementar intervenciones más efectivas en el momento adecuado.

Para validar la efectividad de su propuesta, los autores llevaron a cabo un análisis comparativo entre los mapas de riesgo generados a partir de los datos de redes móviles y los mapas creados con los métodos tradicionales de salud pública. Los informes convencionales, aunque útiles, presentan la limitación de ser más lentos en su capacidad de proporcionar información detallada y actualizada.

En comparación, los mapas de riesgo basados en la movilidad de los usuarios no solo fueron capaces de identificar las áreas de mayor riesgo de manera mucho más rápida, sino que también demostraron ser más precisos en la predicción de la propagación de la epidemia. Al integrar los datos de redes móviles en el análisis de riesgos, los investigadores lograron detectar con mayor antelación las zonas con alta probabilidad de contagio, lo que otorga una ventaja significativa en la implementación de medidas preventivas, como el aislamiento de áreas de alto riesgo o la redirección de recursos hacia zonas específicas.

Este enfoque, por lo tanto, no solo mejora la capacidad de respuesta ante brotes epidémicos, sino que también optimiza la asignación de recursos, permitiendo una respuesta más eficiente y focalizada en los lugares que más lo requieren. En conjunto, los resultados del estudio demuestran que el uso de los datos de redes móviles para monitorear la propagación de epidemias puede transformar la forma en que se gestionan los brotes infecciosos, mejorando los resultados de salud pública y reduciendo el impacto de las enfermedades en la población.

Otro componente esencial de este método es la base de datos de hospitales que permite identificar las señales móviles que se pueden asociar con una alta probabilidad de infección. La base de datos empleada para este proyecto se ha generado a partir de la información pública del ministerio de sanidad disponible sobre los hospitales tanto públicos como privados en las ciudades de Madrid y Barcelona, junto con sus respectivas ubicaciones.

Esta información es fundamental para el análisis, ya que permite identificar los centros hospitalarios a los que podrían haber sido trasladados los usuarios en caso de que fueran hospitalizados durante el periodo del estudio. En este caso, las variables utilizadas en la base de datos de hospitales incluyen el nombre de cada hospital, su ubicación geográfica y, crucialmente, si el hospital estuvo aceptando pacientes durante la pandemia de COVID-19, lo cual se considera un criterio necesario para determinar su relevancia en este estudio.

En resumen, la integración de estas dos bases de datos, la de los usuarios y la de los hospitales, constituye la columna vertebral del análisis realizado. Gracias a la información detallada y actualizada de ambos conjuntos de datos, es posible llevar a cabo un estudio robusto y preciso, que ofrece un análisis sobre los patrones de movilidad de los usuarios en relación con los centros hospitalarios durante la crisis sanitaria provocada por el COVID-19.

En la próxima entrega, como se detalla en el apartado siguiente, se propone dividir la base de datos en distintos nodos para simular el funcionamiento de varias entidades o empresas. Esta estrategia tiene como finalidad modelar cómo se gestionaría el proceso en un contexto de múltiples actores, ofreciendo una perspectiva más completa y detallada de la dinámica que se busca analizar. Este enfoque no solo simula la distribución descentralizada de los datos, sino también las interacciones y la gestión de información entre estas "empresas",

enriqueciendo el análisis y proporcionando una comprensión más profunda del impacto del proceso en escenarios reales.

4. Demostrador

Nuestro caso de uso trata de identificar áreas de riesgo de epidemia con datos de movilidad, para así poder controlar la tasa de contacto. Para ello se han usado los datos de movilidad de la operadora telefónica Orange. El objetivo es calcular zonas de riesgo de contagio asegurando la privacidad de los usuarios y los datos de la entidad colaboradora.

Para este estudio, como ya se ha mencionado anteriormente se han empleado grandes cantidades de datos. Con ellos se han extraído las trayectorias aproximadas de los movimientos de la población. También se han calculado variables de interés como las estimaciones de hospitalizados y el periodo de tiempo de hospitalización.

Se han realizado mapas de calor del flujo de gente a lo largo de la mañana y de la noche de Madrid y Barcelona. Para ello se ha considerado un día concreto y se ha filtrado la variable *"schedule"* para poder representar el flujo de gente a lo largo de la mañana y la noche en cada geohash. Dejando como resultado las siguientes imágenes.

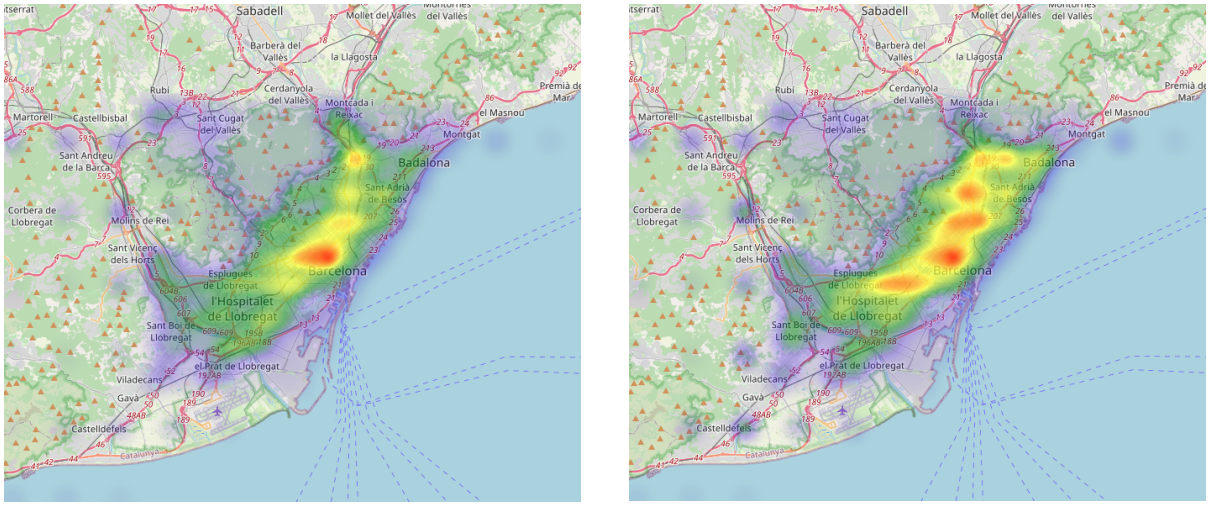

Ilustración 11: Mapas de calor de densidad de población en Barcelona el día 19 de Febrero de 2024 por la mañana (izquierda) y por la tarde (derecha).

Se puede observar que a la mañana la gente se concentra en el centro de Barcelona, mientras que a la noche se distribuye más. Esto corrobora la calidad del dato al confirmar una tendencia evidente, puesto que muchos usuarios se desplazan desde sus viviendas a las afueras de Barcelona para estudiar o trabajar en el centro.

Se puede observar un comportamiento similar en los mapas de calor de Madrid que se muestran a continuación.

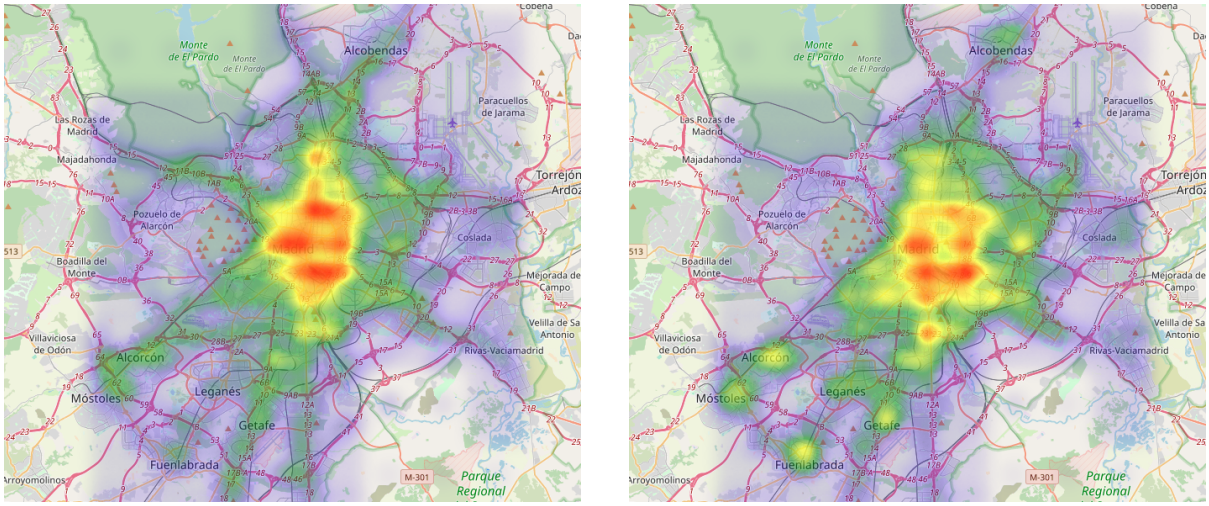

Ilustración 12: Mapas de calor de densidad de población en Madrid el día 19 de Febrero de 2024 por la mañana (izquierda) y por la tarde (derecha).

Adicionalmente, se llevó a cabo un estudio para analizar el número de personas hospitalizadas y determinar la duración de su estancia en el hospital.

En la imagen que se presenta a continuación, se ilustra la distribución del número de personas en función de la cantidad de días consecutivos de hospitalización y el tipo de hospitalización: jornada completa (*Full Day*) o media jornada (*Half Day*). La intensidad del color en cada celda representa la cantidad de personas correspondientes a cada combinación de días consecutivos y tipo de hospitalización. Las celdas más oscuras indican un mayor número de personas, mientras que las más claras reflejan un menor número.

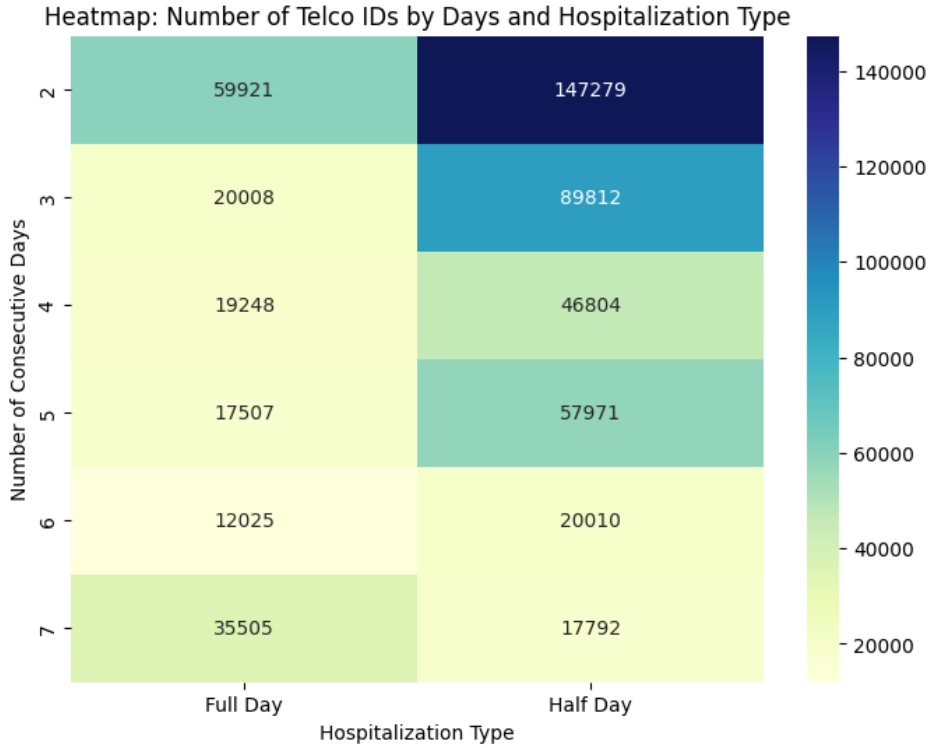

Ilustración 13: Número de dispositivos por día y tipo de hospitalización la semana del 19 de Febrero de 2024.

Para identificar las hospitalizaciones se ha creado la variable "*hospitalization_type*" donde consideramos que el usuario es "*Full Day*" si la variable *"schedule"* (que indica el intervalo del día al que corresponde la actividad del usuario, ya sea mañana, tarde o noche) ha pasado por los tres intervalos posibles. Y se considera "*Half Day*" si tan solo ha pasado por dos de los tres intervalos posibles.

Esto nos permite determinar el número de personas que han estado hospitalizadas durante siete días, así como visualizar el número de personas hospitalizadas por día o que asisten al hospital únicamente medio día.

Recordemos, que este estudio se han eliminado todos aquellos usuarios que viven o trabajan en el hospital, permitiendo así poder centrarnos únicamente en hospitalizados o visitantes.

Más datos, mejor analítica

El demostrador evidencia cómo la incorporación de información más granular, como un mayor número de antenas, y datos de una base ampliada de usuarios permite obtener análisis más precisos. La integración de datos provenientes de antenas de distintas operadoras, junto con los patrones de movilidad de un mayor número de usuarios, proporciona una visión más completa y detallada de las áreas de riesgo durante una pandemia, fortaleciendo la capacidad para tomar decisiones informadas.

5. Conclusión y siguientes pasos

Los estudios realizados para este entregable conforman las dos primeras capas del demostrador diseñado en el Entregable 4.1. En él se establecían tres capas de información:

- Capa de información de los movimientos poblacionales.
- Capa de información de población hospitalizada.
- Capa de información de la población Auto-confinada.

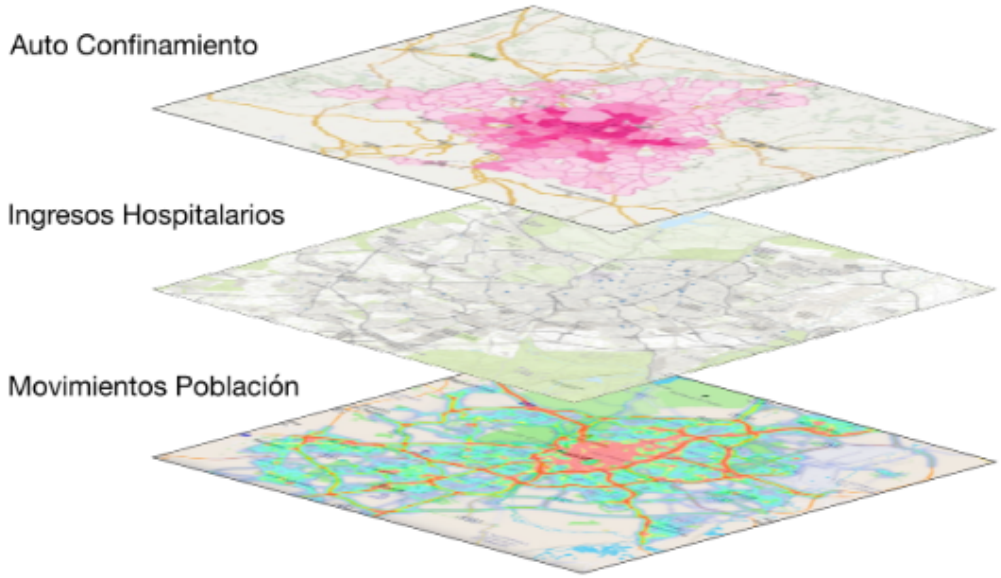

Ilustración 14: Capas de información para el demostrador.

Para ello se han realizado varias tareas imprescindibles para generar las capas del demostrador. Algunas de estas tareas son:

- **Extracción de datos**: Esta tarea ha sido y es un gran reto debido a la volumetría del dato, tal y como se describe en apartados anteriores.
- **Cruce de datos**: Para poder obtener una base de datos con la que trabajar se ha tenido que combinar la información que se disponía de los clientes, las antenas y las geolocalizaciones.
  - **Cálculo de trayectorias**: Al conocer el CGI al que está conectado cada usuario, es posible determinar con precisión el área geográfica en la que se encuentra en un momento dado. Esto no solo permite mapear la ubicación de cada usuario, sino también calcular el tiempo que permanece en cada área, facilitando así el análisis de sus trayectorias y comportamientos de movilidad.
- **Limpieza de datos**: Para poder realizar el estudio se han extraído los datos de las ciudades de interés (Madrid y Barcelona), eliminando datos anómalos como usuarios que recorrían una larga distancia en un tiempo imposible.
- **Creación de base de datos**: Se han creado bases de datos sobre las que trabajar, filtrando las personas que viven o trabajan en un hospital, al igual que aquellas de las que no se dispone dicha información.
- **Búsqueda y modificación de base de datos de hospitales**: Para poder filtrar las personas hospitalizadas es imprescindible saber la ubicación de los hospitales. Para ello se ha empleado una base de datos del ministerio de salud donde figuran los nombres y direcciones de los hospitales de todo España (entre otros datos). Con estos

datos se ha buscado la ubicación en forma de (latitud, longitud) que posteriormente se ha codificado a geohash, dándonos como resultado la base de datos de hospitales que emplearemos a lo largo de todo el estudio.

- **Creación de la capa de movilidad**: Filtrando los datos por día y por *"schedule"* para poder crear un mapa de calor del flujo de gente a lo largo de la mañana, tarde o noche.
- **Creación preliminar de la capa de hospitalizados**: Se ha realizado un mapa de calor determinando el número de personas hospitalizadas durante un periodo de siete días. En él se puede determinar el número de usuarios hospitalizados uno, dos, tres, hasta siete días.

En este entregable se han sentado las bases del demostrador final que incluirá las visualizaciones de las capas ya generadas, además de una tercera con el análisis de la población auto-confinada para tener una visión completa y poder ayudar en la toma de decisiones.

Como parte de los objetivos del proyecto, dentro del próximo entregable, también se iniciará el proceso de federación. Como se mencionó en el entregable anterior, para demostrar la utilidad de las tecnologías federadas en este caso de uso, se dividirán los datos disponibles en dos o más conjuntos disjuntos. El propósito es simular que cada conjunto de datos proviene de un proveedor diferente, para lograrlo se evaluarán tres criterios de selección diferentes:

Generar dos o más conjuntos de identificadores únicos (números de teléfono)
Esto nos permitirá simular la construcción de informes agregados utilizando datos de operadores de telefonía móvil que no tienen clientes en común. Cada uno generará las trayectorias de sus clientes que luego se agregarán en el demostrador mediante tecnologías federadas.

Generar dos o más conjuntos de antenas
Este escenario permitirá simular el procesamiento de datos en el extremo, en la propia antena. Cada proveedor agregará los datos de las antenas de su propiedad y luego podrá utilizar tecnologías federadas para construir las trayectorias finales en conjunto con los datos de otros operadores.

Distribuir los datos en áreas geográficas
En este escenario consideraremos que cada proveedor de datos cubre un área geográfica concreta. De este modo al agregar los datos de varios proveedores con tecnologías federadas podremos cubrir un área geográfica mayor que si utilizaremos los datos de un solo operador.

Dado que el objetivo de este caso de uso es analizar los patrones de movilidad en un área geográfica determinada, el uso de tecnologías federadas permitirá agregar la analítica de varios conjuntos de datos diferentes. Por lo tanto, se utilizarán técnicas como la privacidad diferencial, la k-anonimidad o la computación multi-partita segura para garantizar la confidencialidad de los datos al generar los informes agregados.

El uso de tecnologías federadas para agregar estos datos es clave, ya que las trayectorias generadas son datos especialmente sensibles de los clientes y pueden utilizarse para identificar a individuos sin su consentimiento. Por tanto, es de vital importancia garantizar la privacidad de estos datos cuando se ponen a disposición de terceros o se combinan con otras fuentes de datos.

Las tecnologías federadas son fundamentales para facilitar la colaboración entre competidores dentro de una misma industria, estableciendo un modelo que permite a la industria europea trabajar conjuntamente en la resolución de desafíos clave. Estas tecnologías superan las barreras asociadas a la confidencialidad de los datos, desbloqueando su potencial y permitiendo un enfoque colaborativo para enfrentar problemas reales de manera eficiente y segura. Identificamos dos líneas de trabajo para explorar el potencial de las tecnologías federadas en este caso de uso y que se abordarán en el próximo entregable:

1. **Colaboración entre múltiples operadores** para enriquecer los datos en una zona común donde todos operan o para complementarse mutuamente en áreas donde solo uno o un subconjunto están presentes.
2. **Entrenamiento local del modelo** que colorea los mapas en cada estación base, sin necesidad de transferir todos los datos a una nube centralizada. El aprendizaje federado (FL) permite entrenar el modelo utilizando los datos disponibles en cada estación base. En lugar de enviar grandes volúmenes de datos al lago de datos, solo se transmiten las actualizaciones del modelo local, lo que resulta más escalable y ligero.