Estado del arte y diseño de los componentes del caso de uso de economía digital

MLEDGE - Aprendizaje automático en la nube y en el borde

(Cloud and Edge Machine Learning)

Junio de 2024

Información sobre el entregable

Nombre del documento: Estado del arte y diseño de los componentes 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: P4 - Implementación del caso de uso de economía digital

Tareas: El entregable es resultado del trabajo en los diversos componentes técnicos:
- A4.1: Diseño de los componentes del caso de uso de economía digital

Entregable: E4.1 – Estado del arte y diseño de los componentes de economía digital.

Autores: Francisco José Huidobro Ruiz (Orange), Carmen Reina García (Orange), Alberto Román (Acuratio)

Revisores: Santiago Andrés Azcoitia (IMDEA), Nikolaos Laoutaris (IMDEA)

Historial de Versiones

Versión	Fecha	Resumen de modificaciones
Version 1.0	30-06-2024	Versión inicial del documento

Índice

Introducción

En diciembre de 2022 fue adjudicado a IMDEA Networks el proyecto “MLEDGE - Aprendizaje automático en la nube y en el borde (Cloud and Edge Machine Learning)” (REGAGE22e00052829516, en adelante el ‘Proyecto’ o MLEDGE) por parte del Ministerio de Asuntos Económicos y Transformación Digital del Gobierno de España, con fondos de la Unión Europea dentro del Plan de Recuperación, Transformación y Resiliencia (European Union - NextGenerationEU/PRTR). El proyecto tiene como objetivo habilitar un ecosistema próspero de servicios FL en el borde seguros y eficientes capaces de facilitar el uso de datos personales y B2B confidenciales para entrenar modelos de ML para consumidores mientras se protege la privacidad de los datos y de sus propietarios.

Los **objetivos generales del proyecto** se pueden resumir en los siguientes:

1. Hacer del aprendizaje federado una funcionalidad accesible y de fácil uso en el borde mediante el desarrollo de una capa de software intermedio y componentes que escondan la complejidad del procesamiento y el intercambio de datos que supone.
2. Resolver problemas técnicos asociados al aprendizaje federado en el borde de la nube.
3. Demostrar esta funcionalidad en casos de uso que reflejen problemas reales de la industria que pueden ser resueltos con estas tecnologías.
4. Explotar los resultados del proyecto involucrando a agentes externos y comunicar los hallazgos al público potencial en general.

Uno de los objetivos básicos del proyecto es diseñar, implementar y hacer públicos demostradores que trabajen con datos sensibles de individuos, y alimenten modelos de aprendizaje automático en diferentes campos de la industria. A tal fin, en la primera parte del proyecto se ha realizado una selección de empresas para el desarrollo de la plataforma FLaaS y el monitoreo de costes de computación, así como el diseño e implementación de casos de uso de negocio reales que se beneficien del aprendizaje distribuido en el borde de la nube. La UTE ORANGE – ACURATIO EUROPE, con CIF U70737150 resultó adjudicataria del paquete de trabajo P4 cuyo objetivo es el diseño y la implementación del caso de uso de economía digital.

El presente documento se corresponde con el entregable 4.1 de título “Informe detallado sobre el diseño del caso de uso de economía digital a implementar y cómo utilizará los componentes de MLEDGE” y tiene como objetivo la documentación detallada, incluyendo la necesidad y el problema que requiere el caso de uso, su impacto y potencial explotación, el diseño de su funcionamiento, y las pruebas de sistema.

1. Definición del problema y objetivos

En este apartado se presenta el problema que trata de resolver el caso de uso de economía digital de MLEDGE, comenzando por su contexto. Adicionalmente, se detalla la situación actual y el estado del arte, entendiendo como tal la infraestructura tecnológica existente sobre la cual se apoyará el caso de uso.

1.1. Contexto

Las enfermedades infecciosas se propagan de forma exponencial. La contención es un medio eficaz para frenar la propagación, permitiendo que los sistemas de salud tengan la capacidad de tratar a los infectados. Sin embargo, una contención similar al "confinamiento" altera la productividad de la población, distorsiona la economía (limitando el transporte y el intercambio de productos básicos) y produce miedo y aislamiento social para aquellos que aún no están infectados o que se han recuperado de una infección.

Varias enfermedades infecciosas tienen períodos de incubación y manifestación asintomática, lo que hace complicado su detección y tratamiento eficaz, haciendo difícil la medición del número real de miembros infectados en la población. Pruebas generalizadas para detectar asintomáticos o el rastreo de contactos han demostrado no ser viables. O son muy costosas y no hay recursos suficientes o requieren la colaboración estrecha de la población.

"Las personas en contacto cercano con alguien que está infectado con un virus, [. . .], tienen mayor riesgo de ser infectados ellos mismos y potencialmente infectar aún más a otros. Vigilar de cerca a estos contactos después de la exposición a una persona infectada ayudará a los contactos a recibir atención y tratamiento, y evitará más transmisión del virus".

Organización Mundial de la Salud (OMS)

Durante la última pandemia quedó patente que como sociedad carecíamos de las herramientas y los datos necesarios, para guiar la toma de medidas sanitarias adecuadas en cada momento. Los intentos de pedir la colaboración masiva, por ejemplo, con la instalación de aplicaciones o pidiendo la cumplimentación de formularios, aunque inicialmente prometedores, quedaron deslavazados por la falta de participación y seguimiento de los participantes.

A pesar de las lecciones aprendidas, la memoria se pierde entre crisis sanitarias o con cada cambio de liderazgo y las medidas exitosas se abandonan rápidamente una vez que la crisis ha terminado.

El ejemplo de otros países, como China, sugiere que la monitorización de los patrones de movilidad ayuda a detener la progresión de la infección. Para ello es fundamental reducir el R0 que determina como de contagiosa es una enfermedad infecciosa. R0 es una descripción del número promedio de personas que pueden contraer una enfermedad por contacto con una persona contagiosa. Idealmente, cuanto menor R0 menor será la propagación y más fácil la labor paliativa de los servicios sanitarios.

Los factores que definen R0:

- **Periodo infeccioso:** normalmente viene fijado por la enfermedad.
- **Tasa de contacto:** cuantas personas entran en contacto con una persona contagiosa.
- **Modo de transmisión:** que también viene determinado por la enfermedad.

Por lo que controlar la tasa de contacto es clave para controlar la transmisión de una enfermedad, por lo que tener análisis diarios de los patrones de movimiento de la población española ayudará a las autoridades sanitarias a tomar medidas adecuadas basándose en datos recopilados en tiempo real. Ya que las zonas de riesgo pueden variar de día en día.

La crisis económica derivada de la pandemia de 2020 produjo que el PIB en España cayera un 11,3% en 2020 y que en 2021 solo recuperara un mero 5,5%, por lo que la pandemia en términos económicos costó a la economía española en crecimiento 127.524 millones de euros. Esto supone un coste por habitante de 2.696 euros.

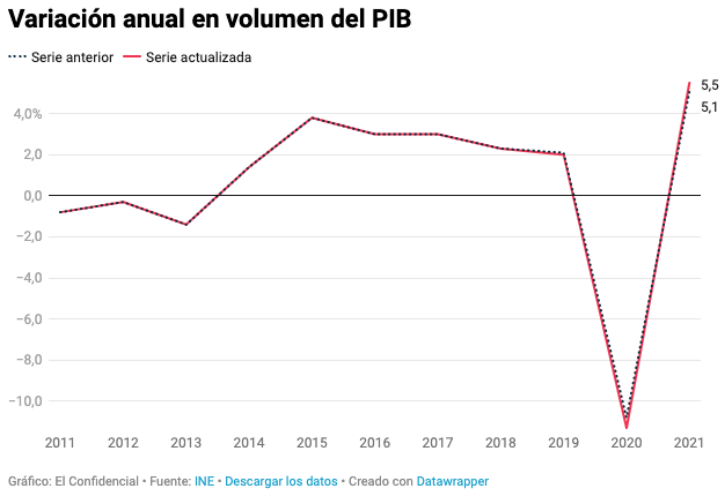


Ilustración 1: Evolución PIB

El impacto económico, sanitario y social es muy alto en cualquier crisis sanitaria. Herramientas que nos ayuden a paliar cualquiera de estos aspectos serán claves en las próximas emergencias.

1.2. 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 contar con el apoyo de todas las operadoras de telefonía con red propia que operan en España, para así tener una imagen completa y en tiempo real del avance de una infección.

Para una primera aproximación se demostrará el potencial de la solución analizando datos de dos grandes ciudades españolas, Madrid y Barcelona. Para ello se obtendrán y procesarán datos de movilidad de los clientes de las operadoras. Estos datos se obtendrán de los Call Detail Records (CDRs) y/o 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 procesamiento, permiten generar trayectorias aproximadas (ubicación y hora) de los usuarios de los dispositivos móviles.

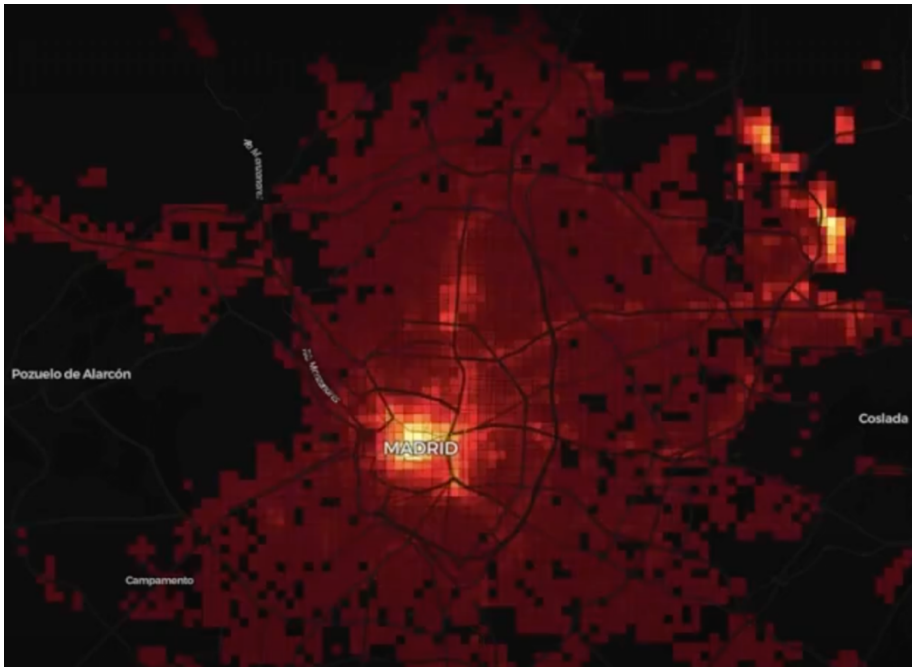


Ilustración 2: Mapa de calos de la movilidad en Madrid

Este análisis dará lugar a la generación de estos mapas de riesgo que podrán 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 confinarse o evitar distintas zonas.

1.3. Situación actual

El dato Telco como solución en casos de uso de Movilidad y Presencia

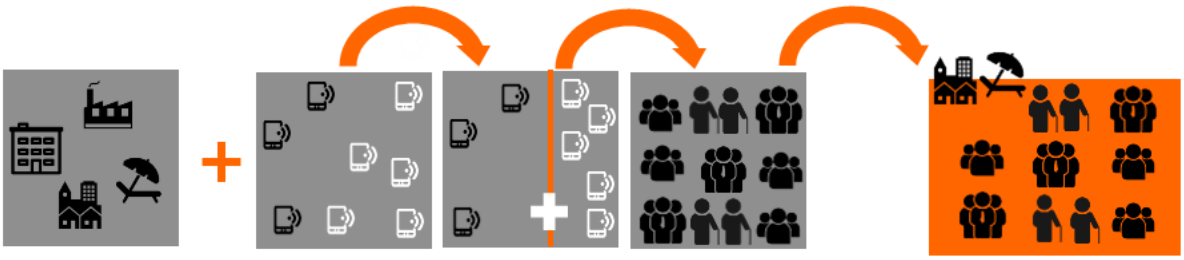
El uso de datos provenientes de la información agregada y anónima de la población a partir de los datos de posicionamiento de los dispositivos móviles es una práctica común y extendida en algunos casos de uso, tanto en el sector privado como en el público. Se trata de la generación de indicadores estadísticos de población basada en el comportamiento anonimizado de los usuarios móviles y su geolocalización en la red móvil del operador en cuestión. Orange viene siendo proveedor de este tipo de soluciones en España en distintos ámbitos desde 2016.

Las soluciones Smart Data de Orange son un conjunto de herramientas analíticas que proveen, entre otros casos, de estadísticas sobre patrones de movilidad y comportamiento de la población; para ello se transforman los registros de señalización de los usuarios móviles anonimizados en indicadores estadísticos, tales como la frecuencia con la que se visitan diferentes áreas geográficas y los desplazamientos de las personas.

La propuesta de valor que este tipo de soluciones aportan a sus clientes incluye:

- **Personalización** de la información adaptándola a las necesidades específicas de cada proyecto en términos de duración y zonas de estudio e indicadores producidos.
- **Fiabilidad y tamaño de la muestra**, ya que consideran todos los usuarios móviles en la red de ORANGE España nacionales e internacionales en itinerancia, una muestra muy superior a la obtenida mediante otros métodos como las encuestas a pie de calle.
- **Optimización** de tiempo y costes en comparación con otros métodos de análisis tradicionales como sondeos y encuestas.
- **Flexibilidad**: acceso a información pasada (histórica) y **recopilación pasiva** de datos.

- **Extrapolación y corrección** de sesgos.



- Conformidad con las normativas europeas de protección de datos (**GDPR**). Ejemplo: E-03690-2020 Resolución de fecha 20-04-2021 Artículo (aepd.es)

La red radio como herramienta de ubicación de usuarios

Las soluciones Smart Data de Orange utilizan el conocimiento profundo de Orange en herramientas analíticas de modelación de cobertura de red radio. La señal de radio entre un terminal móvil y una antena de transmisión es significativamente volátil, por lo que su conocimiento es clave para la ubicación de los usuarios de teléfonos móviles y la generación de indicadores de movilidad fiables.

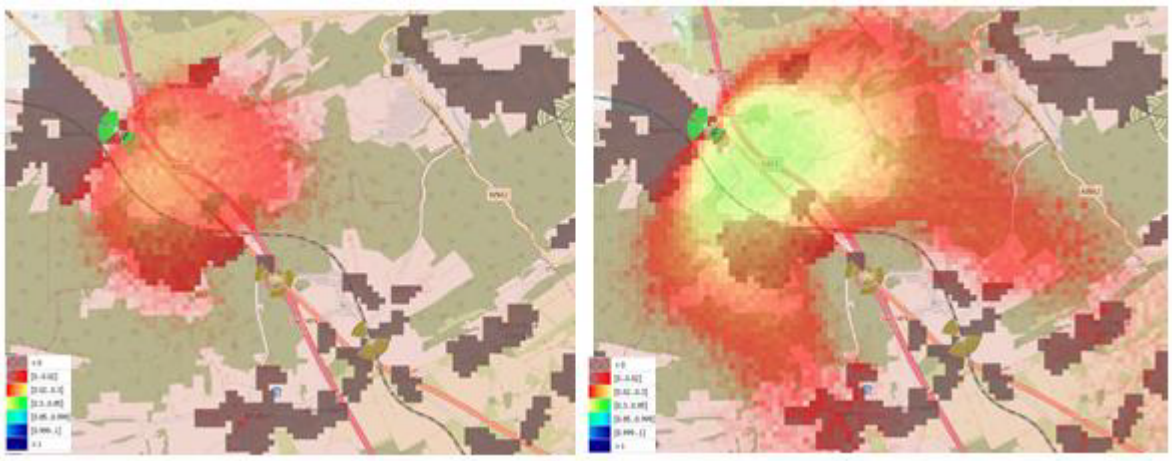


Ilustración 3: Ilustración de la probabilidad de cobertura de la misma antena 4G a 2.6GHz y 800MHz

A continuación, se presentan algunos ejemplos de proyectos relevantes de big data usando datos de empresas de telecomunicaciones:

1.4. Ejemplos de proyectos big data con dato telco relevantes

Entidad: *Ministerio de Transportes y Movilidad Sostenible*

Tipo de Proyecto: *Movilidad*

Título y Expediente: Servicio para la realización del estudio de movilidad de viajeros de ámbito nacional aplicando la tecnología Big Data. Expediente SETMA2020043

Resumen: El objeto del contrato es la realización de un trabajo de consultoría y asistencia técnica necesario para estudiar la movilidad de viajeros a nivel nacional en cada uno de los modos de transporte (carretera, ferrocarril, marítimo, aéreo, etc.), aplicando la tecnología Big

Data. Para ello, se utilizan como fuente principal de datos los que proporcionan los terminales móviles y su señalización en la Red móvil de Orange.

Entidad: *Secretaría de Estado de Transportes, Movilidad y Agenda Urbana*

Tipo de Proyecto: *Movilidad*

Título y Expediente: Análisis de la movilidad en España con tecnología Big Data durante el estado de alarma para la gestión de la crisis del COVID-19. Estudio de movilidad con Big Data durante la pandemia | Informe metodológico | Publicación de los datos abiertos

Resumen: En la elaboración de este análisis se utiliza como fuente principal de datos el posicionamiento de los teléfonos móviles en la red celular, siendo una condición indispensable el cumplimiento de la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales.

Esta metodología, relativamente novedosa y en parte experimental, se consideró como la óptima. Se analizó una muestra de gran tamaño y el proceso de datos y generación de indicadores necesitaba de un corto periodo de respuesta (tres días).

Entidad: *Instituto Nacional de Estadística (INE)*

Tipo de Proyecto: *Movilidad*

Título y Expedientes: Estadística Piloto sobre Movilidad a partir del posicionamiento de teléfonos móviles (Censo de Población y Viviendas 2021).

Expediente: 2019N0060001 ; Expediente: 2019N0060002 ; Expediente: 2019N0060003

Resumen: Este estudio propone el análisis de idoneidad de la información estadística obtenida a partir del posicionamiento de los teléfonos móviles para su incorporación a la información sobre movilidad que se ofrece tradicionalmente en los Censos de Población y Viviendas.

En concreto se solicitaron tres matrices de movilidad con información de los tres principales operadores de telefonía móvil (OTM).

Esta información permitió al INE cotejar con la información obtenida por otros medios analizar su idoneidad. Además, se publicó en la página web del INE dentro del apartado de estadística experimental para así obtener retroalimentación de los usuarios e incluir en el futuro esta fuente de datos dentro del plan de utilización regular de las estadísticas demográficas.

De especial relevancia es este proyecto, en el que se pretende hacer **censo** con información de los 3 principales operadores en ese momento. **La utilización de soluciones como la propuesta en este documento permitiría poder compartir de forma privada y segura los datos de cada operador, unificando criterios en la elaboración de los indicadores y simplificando el proceso.**

1.5. 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” (enlace: https://drive.google.com/file/d/1OQg2dxPu-x-RZzETlpV3lFa259Nrpk1J/view?pli=1) 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 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 proporcionará los siguientes datos:

- De una a dos semanas de datos de movilidad al principio de la pandemia, de los clientes de dos grandes ciudades españolas, como por ejemplo Madrid y Barcelona. Estos datos consisten en las trayectorias a lo largo del día de los clientes, que son una serie de pares de ubicaciones y horas a lo largo del día.
- De una a dos semanas un año después del inicio de la pandemia en las mismas ciudades.
- De una a dos semanas 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 serán utilizados en el Proyecto para generar los entregables descritos a partir de los datos de telefonía, así como los procesos de adecuación necesarios para el mismo:

Para garantizar la privacidad de los usuarios se usan tecnologías habilitadoras de la privacidad. Y en este caso se están probando Federated Analytics de cara a garantizar que ningún usuario de telefonía pueda ser reindentificado y al mismo tiempo obtener la información más precisa de cara ayudar a las administraciones a tomar las mejores decisiones basadas en datos.

CDR y sondas de red.

Se trata de los registros pseudoanonimizados por Orange 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 pseudoanonimizan, 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 en los Dataset descritos en el punto anterior. Además, este inventario recoge la información relevante como las coordenadas del emplazamiento y la información técnica necesaria (potencia de señal, etc.).

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 desplegará su plataforma federada, que tendrá acceso a un repositorio de almacenamiento donde Orange almacenará los datos objeto del estudio. Desde la plataforma se podrán diseñar y entrenar los modelos.

Para alcanzar los objetivos seguiremos tres fases bien diferenciadas:

1. **Extracción de datos tradicional:** Se hará en AWS y consistirá en el procesamiento de eventos de red de dispositivos conectados a antenas de Orange en las áreas geográficas designadas para la reconstrucción de trayectorias.
2. **Conexión a través de la plataforma federada:** Una vez extraídos los datos se volcarán en la nube desde donde el software de la plataforma federada PaaS podrá procesarlos para generar informes agregados que puedan mostrarse en un interfaz unificado.
3. **Implementación de soluciones federadas:** Con los datos cargados en la plataforma federada podrán explorarse desarrollos adicionales que permitan habilitar la posibilidad de compartir los datos de diferentes fuentes (ej. diferentes compañías de

telecomunicaciones) con el objetivo de enriquecer los informes agregados o incluso de construir modelos más complejos.

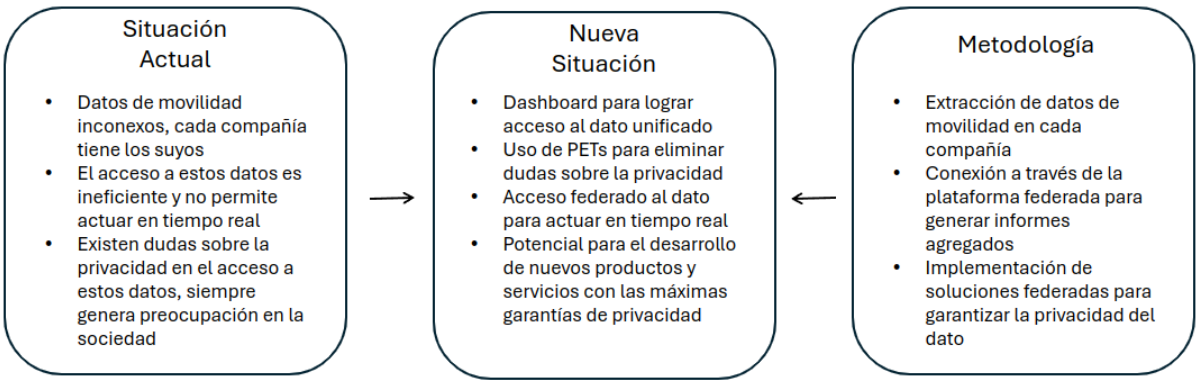


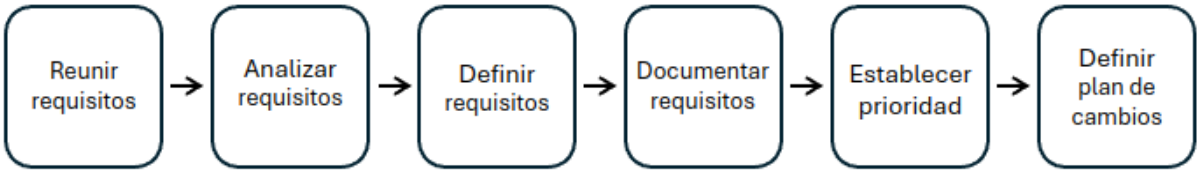
Ilustración 4: Metodología solución

En definitiva, se empleará una metodología tradicional de procesamiento de datos para la extracción de la información relevante a partir de las señales en crudo proporcionadas por Orange. Después, se cargarán los datos en la plataforma federada para mostrar sus capacidades y probar que se puede ofrecer un acceso unificado a datos de movilidad de distintas fuentes preservando la privacidad de los implicados.

2. Especificación de requisitos

2.1. Metodología

Para la elaboración de los requisitos se ha seguido una metodología en seis fases.



Primero, se acuerdan los requisitos de los componentes para el caso de uso propuesto, para que todos los objetivos puedan cumplirse de manera satisfactoria. A continuación, se alinean estos requisitos con los objetivos del caso de uso, y se clasifican en funcionales y no funcionales.

Una vez hecho esto se definen los requisitos de manera que sean claros, precisos y fácilmente comprobables. Este proceso de definición se acompaña de la documentación apropiada para poder alinear cada requisito con su objetivo dentro del caso de uso.

Por último, se establece la prioridad de cada requisito entre opcionales y obligatorios y se establece el procedimiento para incorporar o eliminar requisitos a lo largo del proyecto.

2.2. 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.

En la ilustración 5 se expone esquemáticamente la arquitectura del caso de uso. Tanto los proveedores como los consumidores de datos deberán disponer de credenciales para conectarse a la plataforma FLaaS mediante la cual se construirá el espacio de datos.

Los proveedores harán un procesamiento local de sus datos para limpiarlos y extraer las trayectorias de sus clientes. A continuación, agregarán estas trayectorias utilizando técnicas federadas que permitan anonimizar los datos y prevenir la reidentificación de los clientes. Por último, pondrán los análisis de movilidad agregados a disposición de los consumidores a través de la plataforma federada.

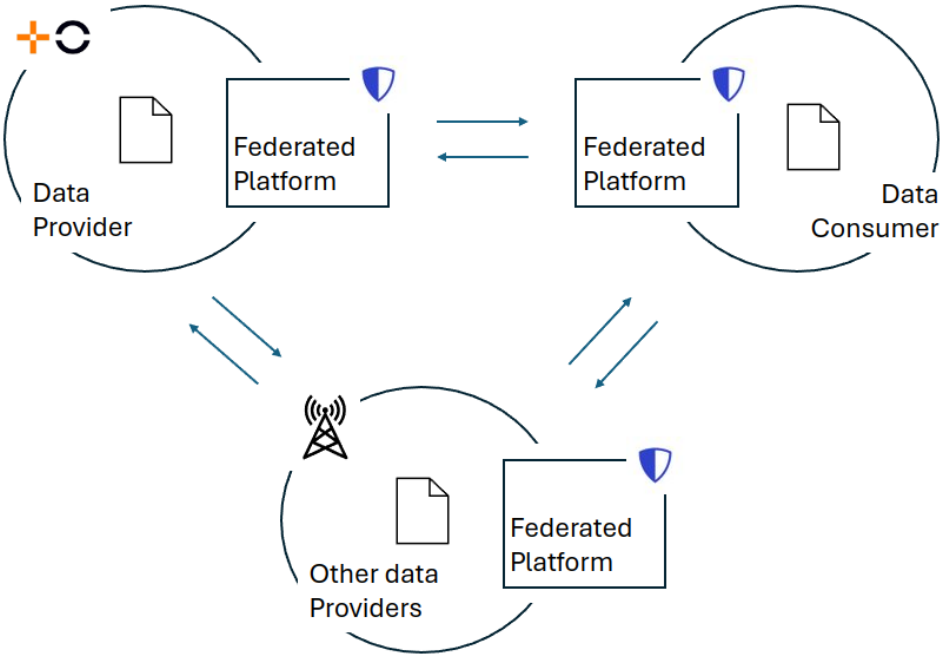


Ilustración 5: Arquitectura del sistema federado

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. Consultar el diseño detallado de los demostradores en la sección 6 para más detalles.

2.3. Requerimientos funcionales

2.3.1. Componente de limpieza y extracción de datos

#	Requisito	Descripción	Tipo
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

#	Requisito	Descripción	Tipo
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
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

2.3.2.Componente de procesamiento federado

#	Requisito	Descripción	Tipo
RF6	Federación	El sistema debe permitir la aplicación de técnicas federadas para la agregación de la analítica de todos los proveedores.	Obligatorio
RF7	Visualización	El sistema debe permitir visualizar la analítica generada para que el usuario final del caso de uso pueda mejorar su proceso de toma de decisiones	Obligatorio
RF8	Transmisión segura	El sistema debe garantizar la seguridad en la transmisión de datos, tanto entre los proveedores como para el consumidor final.	Obligatorio
RF9	Actualización	El sistema debe permitir mantener actualizados los informes agregados, para ello se utilizarán las capacidades de programación de tareas de la plataforma federada.	Opcional
RF9	Procesamiento	El sistema debe permitir procesar los datos mediante un protocolo P2P, en que la información se trasiegue y procese sin terceros de confianza.	Obligatorio

2.4. Requerimientos no funcionales

2.4.1.Componente de limpieza y extracción de datos

#	Requisito	Descripción	Tipo
RNF1	Escalabilidad	Debe permitir el procesamiento de grandes cantidades de datos de forma eficiente	Obligatorio
RNF2	Versatilidad	Debe soportar múltiples formatos de entrada y algoritmos de compresión	Obligatorio
RNF3	Integridad	Debe ser robusto y considerar casos extremos que puedan aparecer en los datos ya que los eventos telefónicos son muy ruidosos	Obligatorio
RNF4	Seguridad	Debe seguir una filosofía de minimización del riesgo, accediendo solo a los datos estrictamente necesarios ya que los ficheros fuente de las señales telefónicas contienen información sensible de los clientes	Obligatorio
RNF5	Adaptabilidad	Debe ser capaz de procesar datos almacenados con ontologías diferentes, o al menos ser fácilmente adaptable para garantizar un procesamiento homogéneo entre datos de distintas compañías	Opcional

2.4.2.Componente de procesamiento federado

#	Requisito	Descripción	Tipo
RNF6	Integración	El envío de datos desde el componente de limpieza y extracción hasta la plataforma federada debe ser sencillo y eficiente	Obligatorio
RNF7	Seguridad	El envío de datos debe ser seguro, dentro de la misma nube o mediante conexiones seguras como Azure Private Link o similares	Obligatorio

#	Requisito	Descripción	Tipo
RNF8	Administración y gestión	Debe existir una interfaz centralizada desde la que controlar quién tiene acceso a los datos subidos a la plataforma y registrar el uso que hace de ellos	Obligatorio
RNF9	Facilidad de Uso	Deben existir interfaces que permitan interactuar con los datos y desplegar procesamientos federados desde el navegador, sin necesidad de conocimientos específicos	Opcional

3. Arquitectura y descripción de los componentes del caso de uso

3.1. Descripción de la arquitectura

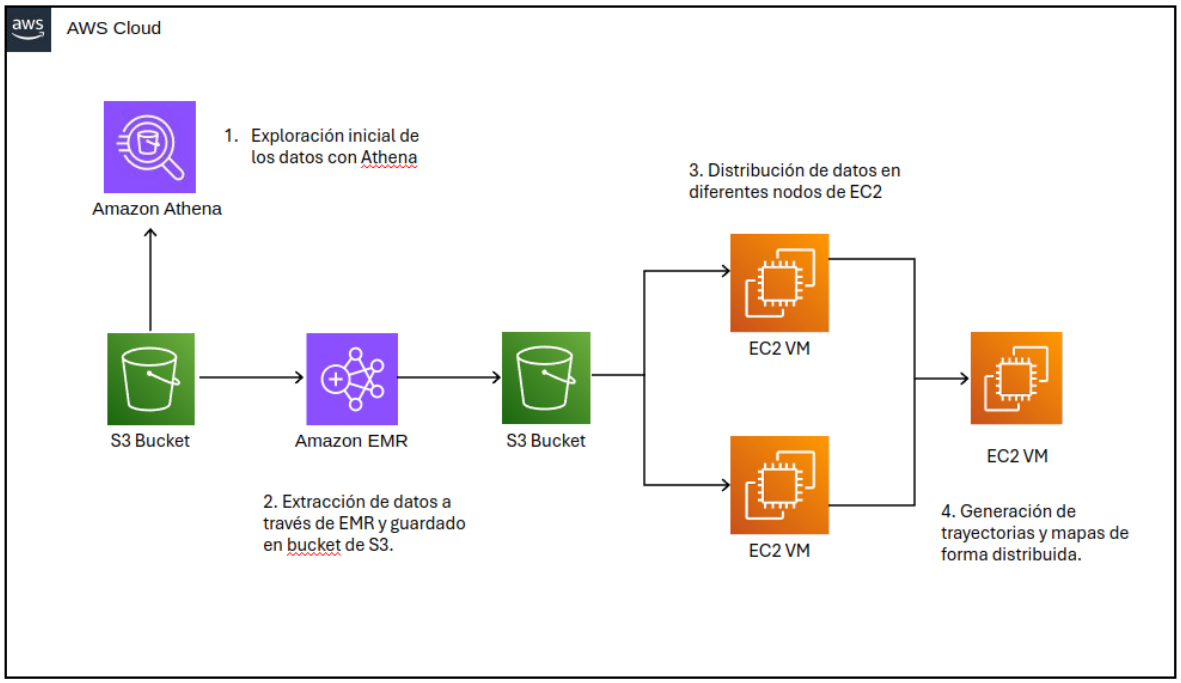


Ilustración 6: Arquitectura adquisición y procesamiento de datos

3.2. Arquitectura software

- **Python:** Este será el lenguaje principal utilizado para el desarrollo de modelos, tanto en entornos locales como federados. Estará preinstalado en el software que Acuratio implementará en las máquinas virtuales EC2.
- **Librería Python Acuratio para secure Multiparty Computation:** El software instalado en las máquinas virtuales EC2 incluirá una librería específica desarrollada en Python por Acuratio para realizar las tareas necesarias para este proyecto.
- **JupyterLab:** Esta será la interfaz que permitirá la interacción con el nodo desde la APP.
- **Aplicación web (APP):** Acuratio desarrollará una aplicación web desde la cual se podrán controlar los nodos.
- **API:** Acuratio proporcionará una API que funcionará como backend de la aplicación web. Esta API actuará como intermediario entre la aplicación y los nodos, así como coordinadora del proceso de entrenamiento.
- **AWS Athena:** herramienta de AWS para facilitar la exploración de bases de datos.
- **AWS EMR**: herramienta de AWS para tratamiento de grandes cantidades de datos de forma distribuida en clusters de computación.

3.3. Arquitectura hardware

- **Bucket S3 de AWS:** contenedor de almacenamiento altamente escalable y duradero en la nube de AWS, diseñado para almacenar una amplia variedad de datos, desde archivos simples hasta grandes conjuntos de datos y contenido multimedia.
- **Máquinas Virtuales** de AWS EC2 con almacenamiento cifrado en EBS: Una máquina virtual por cada operador de telefonía que se vaya a simular y una máquina virtual que actúe de servidor en el entrenamiento federado.

3.4. Matriz de requerimientos - componentes Funcionales

3.4.1.Componente de limpieza y extracción de datos

#	Requisito	Componentes
RF1	Exploración	AWS Athena permitirá la exploración de los datos.
RF2	Preprocesamiento de datos crudos	El procesamiento se realizará con un clúster EMR en AWS.
RF3	Estandarización	Acuratio y Orange definirán las columnas a incluir en el formato estandarizado y el tipo de cada columna.
RF4	Almacenamiento	Los datos se almacenarán en S3, que cuenta con su propio sistema de cifrado.
RF5	Análisis	Acuratio y Orange proveerán herramientas software para realizar el análisis de las trayectorias.

3.4.1.Componente de procesamiento federado

#	Requisito	Componentes
RF6	Federación	El software y las librerías que desarrolle Acuratio permitirán el desarrollo y la federación de modelos.
RF7	Visualización	Los notebooks interactivos de Jupyter Lab integrados en los nodos, así como diferentes menús de la aplicación de Acuratio permitirán a los usuarios del caso de uso total visibilidad durante toda la analítica.
RF8	Transmisión segura	En sus librerías Acuratio implementará medidas de seguridad y encriptación para impedir que ningún nodo o servidor pueda inferir nada de los datos de entrenamiento de otros nodos. Al estar todos los componentes dentro de AWS los datos se transmitirán por su red privada VPC.

#	Requisito	Componentes
RF9	Actualización	La plataforma contará con una forma de programar tareas de forma que se pueda automatizar la actualización de los informes.

3.5. Matriz de requerimientos - componentes No Funcionales

3.5.1. Componente de limpieza y extracción de datos

#	Requisito	Componentes
RNF1	Escalabilidad	Los clústeres de AWS EMR y su control mediante Spark son altamente escalables.
RNF2	Versatilidad	Tanto las librerías Python para análisis de datos como la API de Spark cuentan con herramientas para procesar no solo distintos formatos de entrada de datos sino también toda la variedad de tipos de variables y casos raros o extremos que puedan contener estos.
RNF3	Integridad	
RNF4	Seguridad	Los datos se almacenan cifrados en S3. Los nodos solo accederán a ellos cuando sea necesario. Si los nodos necesitan guardar algo en local, ya sea datos o modelos esto se harán en bloques de almacenamiento cifrados de EBS.
RNF5	Adaptabilidad	El software que Acuratio ofrecerá en su plataforma y nodos se adaptará y podrá trabajar con muchos tipos distintos de datos.

3.5.2. Componente de procesamiento federado

#	Requisito	Componentes
RNF6	Integración	Todas las comunicaciones y transferencias de datos se harán en AWS y sus redes VPC privadas.
RNF7	Seguridad	
RNF8	Administración y gestión	La plataforma de Acuratio llevará integrado un sistema de roles que permitan gestionar que usuarios tienen acceso a los diferentes recursos y datos.
RNF9	Facilidad de Uso	La aplicación contará con interfaces que permitan realizar distintos análisis sin necesidad de ser un experto en la materia. Para niveles más avanzados, las librerías para Python desarrolladas serán lo más parecidas posibles a las ya existentes y ampliamente usadas, de forma que resulten intuitivas.

4. Diseño detallado de la solución

4.1. Limpieza y extracción de datos

Los datos iniciales en crudo se almacenarán en buckets de S3. S3 es el servicio de almacenamiento de objetos en la nube de Amazon Web Services (AWS). Estos datos son grandes volúmenes de registros en formato CSV y comprimidos a formatos gzip y bzip2. Estos ficheros se dividen en 4 tipos principalmente y cada uno se describe a continuación:

- **GEOLOCALIZACIÓN**: Este archivo contiene los datos de geolocalización de los usuarios: id de usuario, CGI, de la antena, timestamp y tipo de evento detectado. Se agruparán los registros cada 5 minutos para reducir la granularidad de los datos y facilitar el análisis de trayectorias. La agrupación temporal ayuda a consolidar eventos cercanos en el tiempo, proporcionando una visión más clara de los movimientos de los usuarios.
- **FOOTPRINT**: Este archivo define las huellas geográficas de las antenas de telecomunicaciones utilizando el formato GeoJSON. Cada registro asocia un CGI (Cell Global Identity) con su correspondiente geometría, lo que permite mapear visualmente las áreas de cobertura de cada antena. Esto es crucial para entender el contexto espacial de los datos de geolocalización.
- **RED**: Este archivo proporciona la información de localización específica de las antenas de telecomunicaciones, incluyendo, CGI, latitud, longitud y código postal. Se tomará el archivo correspondiente al último día de cada semana, ya que estos datos se actualizan diariamente y la versión más reciente ofrece la información más precisa. Además, se deben concatenar los ficheros correspondientes a las distintas tecnologías de antenas (LTE, GSM y UMTS) para obtener una visión integral de la red.
- **CLIENTES**: Este archivo contiene información sobre los clientes, incluyendo su ID único, así como las ubicaciones de su hogar y trabajo. Esta información es esencial para el análisis de patrones de movilidad.

Para proceder a la exploración inicial de estos datos se usará la herramienta Athena de AWS. Athena es un servicio *serverless*, lo que significa que no requiere la configuración ni administración de infraestructura. Es un servicio interactivo de consultas que permite analizar grandes cantidades de datos allí donde estén almacenados, y soportando una gran variedad de tipos de archivos de almacenamiento, utilizando SQL estándar. Los usuarios pueden escribir consultas SQL directamente en la consola de Athena para obtener una comprensión preliminar de los datos, identificar patrones y preparar los datos para análisis más profundos. En nuestro caso, Athena trabaja directamente con los datos almacenados en S3 y facilita la transformación y limpieza de los datos en esta fase inicial.

Una vez hecha una exploración inicial y teniendo una comprensión mejor de los datos, el siguiente paso será hacer una limpieza y extracción a gran escala con el objetivo de generar unos ficheros de trayectorias para los distintos clientes. Este proceso se apoyará en las herramientas Apache Spark a través de un cluster de AWS EMR (Elastic MapReduce).

Apache Spark es un motor de procesamiento de datos distribuido que es conocido por su velocidad y facilidad de uso en el procesamiento de grandes volúmenes de datos. Amazon EMR proporciona un servicio administrado que facilita la ejecución de frameworks de procesamiento de big data, como Apache Spark, en clústeres elásticos y escalables en la nube de AWS.

EMR proporciona acceso a un Notebook de JupyterLab. Los notebooks de JupyterLab son una herramienta interactiva y versátil que permite a los usuarios combinar código, texto,

visualizaciones y ecuaciones en un mismo documento. Esta capacidad interactiva fomenta un flujo de trabajo dinámico. Además, los notebooks de Jupyter soportan múltiples lenguajes de programación, como Python, R y Julia, lo que los convierte en una herramienta flexible para los desarrolladores. Su facilidad para compartir y colaborar hace que sean ideales para equipos que trabajan en proyectos de análisis de datos y desarrollo de modelos de manera conjunta.

En este caso, el trabajo se desarrollará utilizando el lenguaje de programación Python. Se implementará un pipeline de procesamiento de datos en un notebook utilizando la API de Apache Spark para Python (PySpark). Al ejecutar este pipeline, Spark junto con Amazon EMR distribuirán la carga de trabajo entre un clúster de máquinas EC2, lo que acelerará y facilitará la limpieza y extracción de datos de la gran cantidad de ficheros disponibles. Esta configuración aprovecha la escalabilidad y potencia de Spark y EMR para manejar eficientemente el procesamiento masivo de datos.

El producto final de este procesamiento serán unos archivos CSV que contendrán, como mínimo, las siguientes columnas calculadas para las áreas geograficas definidas dentro de Madrid y Barcelona:

- **telco_id:** MSIDSN hasheado (número de teléfono)
- **geohash_id:** Identificador estandarizado del área geográfica en la que se ha registrado el evento asociado al telco_id.
- **month:** Mes en el que se ha registrado el evento.
- **day:** Día en el que se ha registrado el evento.
- **schedule:** Momento del día (mañana, tarde, noche)
- **elapsed_time:** Duración del evento.
- **motor:** Si el cliente iba a pie o en un vehículo.
- **id_fecha:** Fecha completa en la que se ha registrado el evento.

A partir de estos datos podrán construirse fácilmente las trayectorias de los clientes para cumplir con los objetivos del caso de uso.

4.2. Federated Computation

Los tres elementos que se quieren estudiar con este proyecto son:

- Mapas con los que visualizar trayectorias.
- Identificar pernoctas en hospitales de personas que no trabajan allí.
- Identificar auto confinamientos domiciliarios.

Se pretende demostrar que este estudio se podría hacer combinando los datos de diferentes compañías telefónicas de manera privada. Para ello se dividirán los datos en dos o tres máquinas distintas de forma que se pueda simular este entorno en el que varias compañías colaboran en un entorno distribuido.

Para poder generar estos mapas de manera distribuida, Acuratio desarrollará un software que implemente protocolos de agregación segura. Un Protocolo de Agregación Segura (PAS) es una técnica que garantiza que el servidor que realiza la agregación no pueda aprender nada de los datos recibidos del resto de nodos.

Esto se logra encriptando estos datos utilizando protocolos de Computación Segura entre Partes. Así, los datos encriptados son irreconocibles para el servidor, pero el resultado de la agregación sigue siendo el mismo que si las actualizaciones estuvieran desencriptadas.

Un requerimiento técnico que tendrá esta solución es el requerir de un tercer servidor en caso de que los datos se hayan repartido solamente en dos nodos. En este caso si el servidor recibe los datos encriptados del otro nodo, encripta los suyos y agrega ambos, puede conocer los datos del otro nodo desagregando del resultado sus datos sin encriptar. En el caso de tres o más nodos este requerimiento va perdiendo importancia, pero en todo caso puede ser interesante siempre en un entorno federado en el que los datos pertenecen a distintas compañías que sea un servidor de un tercero quien realice la agregación y no el nodo de una de las compañías participantes.

Los nodos de computación, así como el servidor en caso de usarlo de forma independiente serán máquinas virtuales de EC2. Estás máquinas tendrán instalado un software de Acuratio capaz de realizar las computaciones explicadas anteriormente. La conexión con los nodos se realizará a través de la Aplicación web desarrollada por Acuratio y de un Notebook de JupyterLab en cada nodo. Esto permitirá crear nodos y poder ejecutar código de Python sobre los nodos.

El desarrollo de ese software, así como la generación de mapas y trayectorias se hará enteramente en Python programando sobre notebooks de JupyterLab.

Por un lado, Python es el lenguaje de programación elegido porque es versátil que se puede utilizar para varias tareas, desde la limpieza y preparación de datos hasta el desarrollo de modelos de Machine Learning, pasando por lo que más puede interesar para este proyecto: potentes herramientas de visualización de datos que permitirán crear todo tipo de mapas.

Además, al ser uno de los lenguajes más usados en análisis de datos, cuenta con una amplia comunidad de desarrolladores que han creado una amplia gama de bibliotecas especializadas en este sector, como Pandas, NumPy, SciPy, Scikit-learn, Matplotlib y Seaborn. Estas bibliotecas ofrecen herramientas poderosas y eficientes para manipular datos, realizar cálculos estadísticos, visualizar información y desarrollar modelos de aprendizaje automático o visualizar datos y gráficas. Los nodos de Acuratio tendrán incorporadas todas estas librerías. Además, la amplia comunidad con la que cuenta Python también ayuda a que siempre haya ayuda disponible on-line cuando se busca información sobre desafíos técnicos o se necesita orientación en el análisis de datos.

5. Diseño detallado de los demostradores

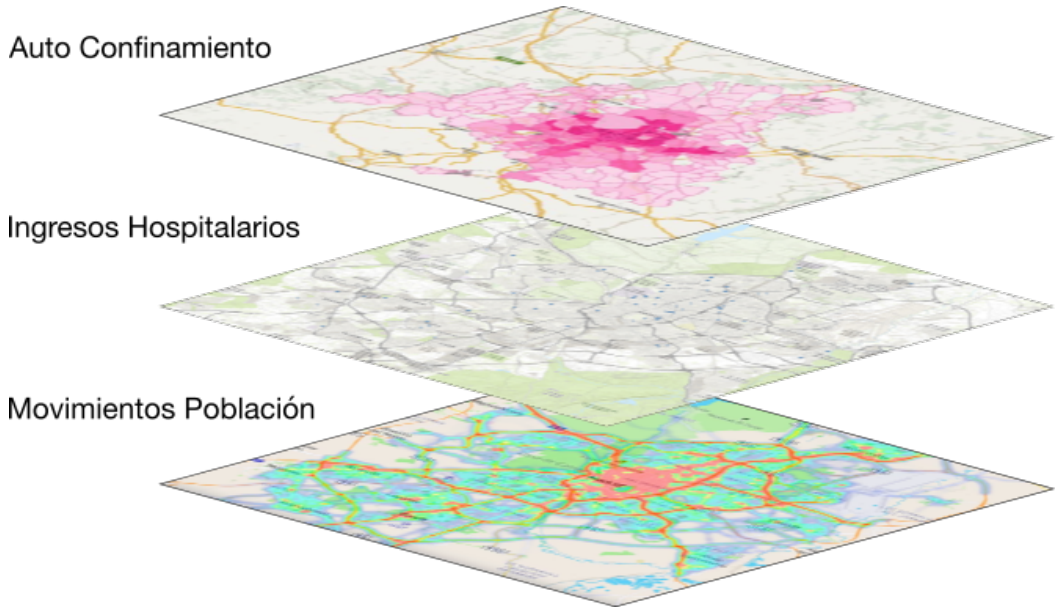
Para controlar la tasa de contacto, nuestro caso de uso tratará de identificar áreas de riesgo de epidemia con datos de movilidad de operadoras telefónicas. Se calcularán zonas de riesgo de contagio, asegurando la privacidad de los individuos y los datos de cada una de las entidades colaboradoras.

Se procesarán grandes cantidades de datos para extraer las trayectorias aproximadas de los movimientos poblacionales, se pseudo-anonimizarán y anonimizarán los datos y se calcularán variables de interés, como las estimaciones de hospitalizados por contagio y los auto-confinamientos.

Para explorar los datos se creará una interfaz en la que se muestren las distintas capas de información a agregar de manera segura y privada con Analítica Federada.

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

Para representar esta información se crearán mapas de calor con los datos agregados de forma federada.



Con esta información también se estimará la tasa actual y la esperada de ocupación hospitalaria para cada área de intereses. De este modo las administraciones públicas dispondrán de una herramienta para la planificación de emergencias o para la asignación de recursos en función de las ratios de contagio previstas.

Se proporcionarán graficas interactivas de estas interfaces para poder explorar de forma dinámica los datos. Por ejemplo, los mapas de calor representarán la situación en distintos momentos de día y se podrá ver la evolución a lo largo del tiempo.

Para ilustrar el proceso completo de procesamiento, agregación y carga en el panel de visualización, se disponibizarán los nodos de la plataforma como sandbox para poder ejecutar distintos procesamientos en tiempo real. Asimismo, se preparará un vídeo de demostración que ilustre el proceso y sirva como guía.

Para demostrar la necesidad de 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 no es construir un modelo, sino 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 multipartita segura para garantizar la confidencialidad de los datos al generar los informes agregados.

La gran ventaja de la agregación mediante tecnologías federadas consiste en la posibilidad de acceder a análisis de movilidad de un conjunto mucho mayor de datos (clientes de varias operadoras vs la cuota de mercado de solo una de ellas), sin comprometer la privacidad de estos clientes. Dado que los datos de movilidad pueden utilizarse para identificar a individuos, el uso de técnicas de privacidad como las tecnologías federadas es esencial para este caso de uso.

El uso de tecnologías federadas para agregar estos datos es necesario, ya que las trayectorias individuales de los clientes pueden utilizarse para identificar a individuos. Por tanto, hay que ser muy cauteloso al combinar estos datos con otras fuentes.

En cuanto a los usuarios de este caso de uso, podemos definir dos actores principales: los proveedores de datos, y los consumidores de la analítica. Ambos interactuarán con los datos a través de la plataforma federada:

- Los proveedores de datos cargarán las tablas procesadas sobre las señales de sus clientes en la plataforma, las agregarán con las de otros proveedores utilizando técnicas federadas y las disponibilizarán también a través de la plataforma. Este sistema permite que los datos nunca salgan de la plataforma federada, donde cada propietario de datos tiene la capacidad de controlar quién y en qué condiciones accede a su analítica.
- Los consumidores de datos accederán a través de la plataforma a las visualizaciones generadas con las tablas agregadas, solo aquellos consumidores autorizados por los propietarios de los datos podrán acceder a las visualizaciones garantizando de este modo la adecuada gobernanza del dato.