NOMMON	**Informe técnico**	Ref.	TR-TTP-20009
		Versión	1
		Date	27/03/2020

Título

Repositorio de datos

Palabras clave

Datos, repositorio, inventario de datos, tarjeta transporte público, validaciones, red de transporte público, GTFS, paradas, estaciones, viajes, usos del suelo, telefonía móvil.

Resumen

El presente documento describe la plataforma utilizada como repositorio de los datos de entrada de los algoritmos para la generación de indicadores de movilidad en transporte público y el sistema de acceso y de licencias de lectura y escritura, así como realizar un inventario de los datos almacenados en el repositorio.

Elaborado por	Javier Monroig	Fecha: 23 de marzo de 2020
Revisado por	David Toribio	Fecha: 26 de marzo de 2020
Aprobado por	Miguel Picornell	Fecha: 27 de marzo de 2020

Distribución

Entidades colaboradoras del proyecto

Registro de Revisiones

Versión	Fecha	Modificación	Secciones afectadas
1	27/03/2020	Versión inicial	N/A

Tabla de contenidos

1. Introducción

1.1 Contexto y motivación

El presente documento se enmarca en el proyecto de investigación industrial **BigData4PublicTransport**, financiado por el Ministerio de Asuntos Económicos y Transformación Digital que pretende desarrollar una nueva tecnología capaz de procesar datos de sistemas inteligentes de pago y combinarlos con otras fuentes de datos (posicionamiento de vehículos, usos del suelo, datos anonimizados de telefonía móvil, etc.), para analizar los patrones de comportamiento de usuarios de transporte público y generar indicadores de movilidad para la planificación y gestión eficiente de los sistemas de transporte público.

1.2 Objeto y alcance del documento

Los datos de entrada de la solución propuesta para la generación de indicadores de movilidad en transporte público se han almacenado en un repositorio de datos con el objetivo de almacenar y compartir la información con la entidad financiadora y las entidades colaboradoras del proyecto. El objetivo principal de este documento es especificar la plataforma utilizada como repositorio de datos; describir el sistema de licencias de acceso, lectura y escritura de los usuarios; y así como realizar un inventario de los datos almacenados en el repositorio.

1.3 Documentos de referencia

En la fecha de realización del presente documento, se han redactado los siguientes informes:

- **Análisis del estado del arte**, donde se realiza una evaluación de la bibliografía científica relacionada con la obtención de información de demanda de transporte público a partir de datos de sistemas inteligentes de pago.
- **Necesidades y requisitos**, donde se detallan las necesidades de información, incluyendo los indicadores específicos de movilidad, y otros requisitos funcionales planteados por el grupo de colaboradores del proyecto.
- **Informe de evaluación de datos**, donde se identifican, caracterizan y evalúan los datos relevantes para la ejecución del proyecto que se encuentran disponibles.

1.4 Estructura del documento

Este documento se estructura en las siguientes secciones:

- **Repositorio de datos**, donde se presenta la tecnología utilizada para el repositorio de datos, así como los permisos de acceso, lectura y escritura del repositorio.
- **Inventario de datos**, donde se catalogan los datos disponibles en el repositorio, así como la estructura de carpetas en la que se organizan.

2. Repositorio de datos

Durante la ejecución del proyecto *BigData4PublicTransport* se ha desplegado un repositorio de datos con el objetivo de servir de plataforma de almacenamiento y transferencia de los datos utilizados durante el proyecto. El servicio desplegado se basa en la implementación de UNIX del protocolo SFTP (SSH File Transfer Protocol), concretamente la implementada por OpenSSH en la distribución de Linux Ubuntu Server LTS.

El repositorio de datos se ha desplegado de forma que los usuarios puedan tener uno de los siguientes roles:

1. **Administrador**: se cuenta con los permisos necesarios para realizar todas las acciones definidas por el modelo de persistencia CRUD (Create, Read, Update y Delete) en cualquiera de los ficheros y carpetas existentes en el repositorio.
2. **Lector**: solo se tienen permisos para listar los ficheros y directorios contenidos en el repositorio, así como leer sus metadatos: tamaño, fecha de última modificación, propietario y permisos. No se cuenta con permisos para acceder al contenido de los ficheros.

El repositorio de datos se ha implementado dentro de la subred DMZ de la infraestructura técnica de Nommon. El servicio se encuentra disponible en la dirección IP pública “85.62.54.200” sobre el puerto 22. El acceso a este servicio se encuentra protegido por un firewall que solo permite conexiones entrantes procedentes de direcciones IP conocidas, por lo tanto, es necesario dar de alta las direcciones IP públicas de los usuarios que vayan a hacer uso del servicio fuera de la red de Nommon.

El repositorio de datos implementa un mecanismo de autenticación que utiliza pares nombre de usuario / contraseña para verificar la identidad de los usuarios. También cuenta con un mecanismo de autorización que permite gestionar los roles descritos anteriormente.

Todo el tráfico entrante/saliente se cifra haciendo uso de protocolos de seguros basados en SSL (Secure Sockets Layer), para que los usuarios puedan verificar la identidad del servicio de repositorio de datos y tener garantía de la autenticidad, integridad y privacidad de los datos transferidos.

El código fuente del software OpenSSH que utiliza el repositorio de datos está disponible bajo licencia de código abierto BSD (Berkeley Software Distribution), lo que permite una rápida detección y corrección de los potenciales problemas de seguridad por parte de la comunidad.

3. Inventario de datos

Esta sección incluye un inventario de los datos disponibles para el proyecto *BigDataforPublicTransport*, así como la estructura de carpetas en la que se organizan dichos datos en el repositorio.

3.1 Datos de la red y oferta de transporte público

Los datos de la red y la oferta de transporte público se recogen en su formato estándar dentro de la carpeta '*/repository/public_transport_network*'. En este directorio se creará una carpeta específica para cada caso de estudio, en la que se incluirán ficheros en formato '.txt' con información estandarizada de la red y la oferta de transporte público, replicando el formato GTFS[1]. En particular, se incluyen los ficheros de:

- agencia (*agency.txt*)
- rutas de transporte público (*routes.txt*)
- paradas de transporte público (*stops.txt*)
- expediciones (*trips.txt*)
- secuencias de parada y horas de llegada de cada expedición (*stop_times.txt*)
- recorrido de las rutas (*shapes.txt*)
- programa (*calendar.txt*)

3.2 Datos de validaciones de tarjeta de transporte público

Los datos de validaciones realizadas con sistemas inteligentes de pago se ubican en el directorio '*/repository/smartcard_validations*', donde se creará una carpeta específica para cada caso de estudio. En ella se incluirán los ficheros de validaciones diarias en formato '.json.gz' organizados en una estructura de carpetas por año, mes y día, como por ejemplo:

/repository/smartcard_validations/madrid/2019/03/17/std_validations.json.gz

Los ficheros estandarizados de validaciones que se subirán al repositorio cuentan con los siguientes campos:

- Identificador de agente (*agent_id*)
- Fecha y hora (*timestamp*)
- Código de operador del servicio donde se realiza la validación (*operator_id*)
- Código de parada donde se realiza la validación (*stop_id*)
- Código del vestíbulo por el que se accede a la estación (*lobby*)
- vehículo en el que se produce la validación (*vehicle_id*)
- Booleano que indica si la validación es de entrada o salida de la red (*exit*)
- Booleano que indica si la validación es inválida (*invalid*)

[1] https://developers.google.com/transit/gtfs/reference

3.3 Datos de usos del suelo

Los datos de usos del suelo obtenidos del Sistema de Información sobre Ocupación del Suelo de España (SIOSE), se han trasladado a una malla compuesta por celdas de 125m*125m y que cubre la totalidad del territorio nacional con el objetivo de agilizar los cálculos. Se subirán los datos de los distintos usos del suelo asociados al código de cada celda de la malla en formato '.txt.gz' a la siguiente carpeta:

'*/repository/land_use*'

3.4 Datos sociodemográficos

Se subirán al repositorio los datos de población por sección censal, género y franja etaria (en grupos de 5 años) obtenidos del Instituto Nacional de Estadística (INE), así como datos de renta promedio por sección censal obtenidos del Atlas de Distribución de la Renta. Los ficheros correspondientes se ubicarán en la siguiente carpeta:

'*/repository/sociodemographics*'

3.5 Matrices de telefonía móvil

Nommon genera una serie de indicadores de presencia y movilidad de la población a partir del análisis de datos de telefonía móvil y su fusión con otros datos (por ejemplo, sociodemográficos, usos del suelo, oferta de transporte, etc.):

- matrices de viajes origen-destino
- presencia de población

Esta información anonimizada se puede generar a un nivel de desagregación espacial de secciones censales o agregaciones de las mismas. Adicionalmente, estos datos se pueden segmentar por:

- lugar de residencia (sección censal),
- tipo de actividad realizada,
- sexo,
- edad,
- nacionalidad
- renta promedio del lugar de residencia
- modo de transporte en el caso de matrices de viajes

Los ficheros con los indicadores de movilidad generados con datos de telefonía móvil se subirán en formato '.txt.gz' en la siguiente carpeta:

'*/repository/mobilephone_matrices*'

3.6 Zonificación de estudio

La solución ofrecerá la posibilidad de generar indicadores de transporte público para una zonificación arbitraria. En caso de no incluir una zonificación específica, por defecto se generarán los indicadores de movilidad para una malla compuesta por celdas de 125m*125m. Cada caso de estudio contará con una carpeta específica con la zonificación especificada por la entidad colaboradora respectiva, por ejemplo:

'*/repository/study_zoning/madrid*'