GRADO EN INGENIERÍA EN TECNOLOGÍAS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

INGESTA DE DATOS EN UN PROYECTO BIG DATA

Autor: Beltrán Rodríguez-Mon Barrera

Director: David Contreras Bárcena

Madrid

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título

INGESTA DE DATOS EN UN

PROYECTO BIG DATA

en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el curso académico 2018/19 es de mi autoría, original e inédito y no ha sido presentado con anterioridad a otros efectos.

El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido tomada de otros documentos está debidamente referenciada.

Fdo.: Beltrán Rodríguez-Mon Barrera Fecha: 12 / 6 / 2019

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

Fdo.: David Contreras Bárcena Fecha: 12 / 6 / 2019

AUTORIZACIÓN PARA LA DIGITALIZACIÓN, DEPÓSITO Y DIVULGACIÓN EN RED DE PROYECTOS FIN DE GRADO, FIN DE MÁSTER, TESINAS O MEMORIAS DE BACHILLERATO

1º. Declaración de la autoría y acreditación de la misma.

El autor D. Beltrán Rodríguez-Mon Barrera

DECLARA ser el titular de los derechos de propiedad intelectual de la obra: Ingesta de Datos en un Proyecto Big Data, que ésta es una obra original, y que ostenta la condición de autor en el sentido que otorga la Ley de Propiedad Intelectual.

2º. Objeto y fines de la cesión.

Con el fin de dar la máxima difusión a la obra citada a través del Repositorio institucional de la Universidad, el autor **CEDE** a la Universidad Pontificia Comillas, de forma gratuita y no exclusiva, por el máximo plazo legal y con ámbito universal, los derechos de digitalización, de archivo, de reproducción, de distribución y de comunicación pública, incluido el derecho de puesta a disposición electrónica, tal y como se describen en la Ley de Propiedad Intelectual. El derecho de transformación se cede a los únicos efectos de lo dispuesto en la letra a) del apartado siguiente.

3º. Condiciones de la cesión y acceso

Sin perjuicio de la titularidad de la obra, que sigue correspondiendo a su autor, la cesión de derechos contemplada en esta licencia habilita para:

a) Transformarla con el fin de adaptarla a cualquier tecnología que permita incorporarla a internet y hacerla accesible; incorporar metadatos para realizar el registro de la obra e incorporar "marcas de agua" o cualquier otro sistema de seguridad o de protección.
b) Reproducirla en un soporte digital para su incorporación a una base de datos electrónica, incluyendo el derecho de reproducir y almacenar la obra en servidores, a los efectos de garantizar su seguridad, conservación y preservar el formato.
c) Comunicarla, por defecto, a través de un archivo institucional abierto, accesible de modo libre y gratuito a través de internet.
d) Cualquier otra forma de acceso (restringido, embargado, cerrado) deberá solicitarse expresamente y obedecer a causas justificadas.
e) Asignar por defecto a estos trabajos una licencia Creative Commons.
f) Asignar por defecto a estos trabajos un HANDLE (URL *persistente).*

4º. Derechos del autor.

El autor, en tanto que titular de una obra tiene derecho a:

a) Que la Universidad identifique claramente su nombre como autor de la misma
b) Comunicar y dar publicidad a la obra en la versión que ceda y en otras posteriores a través de cualquier medio.
c) Solicitar la retirada de la obra del repositorio por causa justificada.
d) Recibir notificación fehaciente de cualquier reclamación que puedan formular terceras personas en relación con la obra y, en particular, de reclamaciones relativas a los derechos de propiedad intelectual sobre ella.

5º. Deberes del autor.

El autor se compromete a:

a) Garantizar que el compromiso que adquiere mediante el presente escrito no infringe ningún derecho de terceros, ya sean de propiedad industrial, intelectual o cualquier otro.
b) Garantizar que el contenido de las obras no atenta contra los derechos al honor, a la intimidad y a la imagen de terceros.
c) Asumir toda reclamación o responsabilidad, incluyendo las indemnizaciones por daños, que pudieran ejercitarse contra la Universidad por terceros que vieran infringidos sus derechos e intereses a causa de la cesión.

d) Asumir la responsabilidad en el caso de que las instituciones fueran condenadas por infracción de derechos derivada de las obras objeto de la cesión.

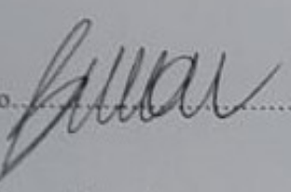
6º. *Fines y funcionamiento del Repositorio Institucional.*

La obra se pondrá a disposición de los usuarios para que hagan de ella un uso justo y respetuoso con los derechos del autor, según lo permitido por la legislación aplicable, y con fines de estudio, investigación, o cualquier otro fin lícito. Con dicha finalidad, la Universidad asume los siguientes deberes y se reserva las siguientes facultades:

- La Universidad informará a los usuarios del archivo sobre los usos permitidos, y no garantiza ni asume responsabilidad alguna por otras formas en que los usuarios hagan un uso posterior de las obras no conforme con la legislación vigente. El uso posterior, más allá de la copia privada, requerirá que se cite la fuente y se reconozca la autoría, que no se obtenga beneficio comercial, y que no se realicen obras derivadas.
- La Universidad no revisará el contenido de las obras, que en todo caso permanecerá bajo la responsabilidad exclusive del autor y no estará obligada a ejercitar acciones legales en nombre del autor en el supuesto de infracciones a derechos de propiedad intelectual derivados del depósito y archivo de las obras. El autor renuncia a cualquier reclamación frente a la Universidad por las formas no ajustadas a la legislación vigente en que los usuarios hagan uso de las obras.
- La Universidad adoptará las medidas necesarias para la preservación de la obra en un futuro.
- La Universidad se reserva la facultad de retirar la obra, previa notificación al autor, en supuestos suficientemente justificados, o en caso de reclamaciones de terceros.

Madrid, a 12 de Junio de 2019

ACEPTA



Fdo. .....

Motivos para solicitar el acceso restringido, cerrado o embargado del trabajo en el Repositorio Institucional:

GRADO EN INGENIERÍA EN TECNOLOGÍAS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

INGESTA DE DATOS EN UN PROYECTO BIG DATA

Autor: Beltrán Rodríguez-Mon Barrera

Director: David Contreras Bárcena

Madrid

Agradecimientos

A mis compañeros de proyecto Cayetano y Pablo, y a mi director David, por ayudarme en todo lo necesario, y porque juntos hemos conseguido realizar este gran proyecto.

A mis padres y a mis hermanos, por aguantarme durante este año.

A mis amigos. Por todo.

INGESTA DE DATOS EN UN PROYECTO BIG DATA

Autor: Rodríguez-Mon Barrera, Beltrán.
Director: Contreras Bárcena, David.
Entidad Colaboradora: ICAI – Universidad Pontificia Comillas

RESUMEN DEL PROYECTO

Palabras clave: API, WebScrapping, Big Data, Raw Data

1. Introducción

La informatización de una gran cantidad de elementos utilizados en el día a día de las personas está provocando que se generen multitud de datos. Esta información puede ser procesada para acabar generando diversas tomas de decisiones según el objetivo final enfocado en un proyecto.

Es de una gran importancia proveer a un sistema Big Data de una diversidad de datos que se complementen entre sí para nutrir a un sistema de decisión de las herramientas suficientes para que este pueda tomar la mejor resolución posible.

2. Definición del proyecto

Este proyecto trata de la elaboración de un sistema de ingesta de datos para su posterior integración y procesamiento en una infraestructura Big Data. La ingesta de datos se realizará mediante tecnologías de captación de datos tales como el WebScrapping, el acceso a APIs en tiempo real u otros tipos de ingesta de datos multifuente.

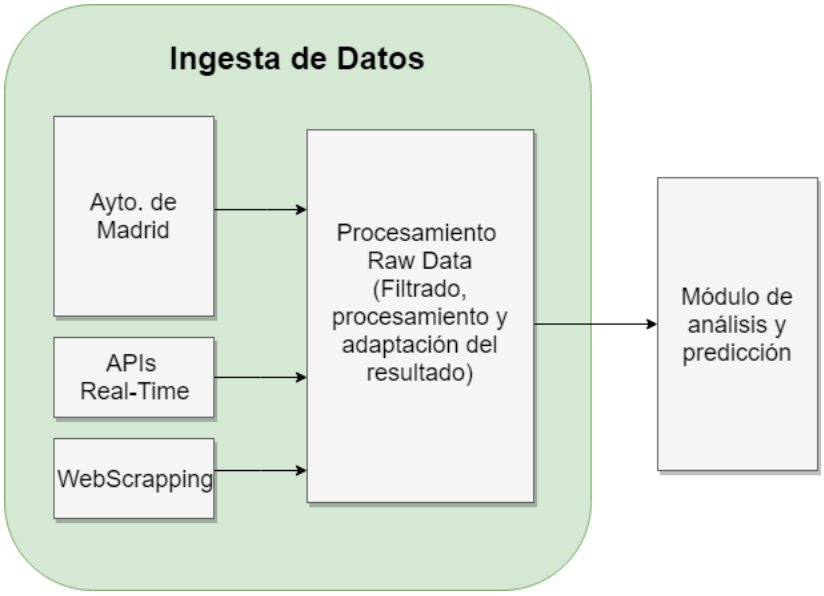
La fuente principal de datos será el repositorio abierto proporcionado por el ayuntamiento de Madrid, focalizando los datos a complementar un set de trayectos recogidos durante varios meses de los años 2018 y 2019.

3. Descripción del modelo/sistema/herramienta

La captación de datos se realizará de forma manual de diversas fuentes (el ayuntamiento de Madrid principalmente, pero también se adaptarán datos de otras fuentes para el sistema).

Tras esto se realizará una limpieza de datos en un clúster Big Data ofrecido por la Universidad ICAI, reduciendo los datos a únicamente la información necesaria para etapas posteriores del Proyecto, y adaptándolos para que su posterior proceso sea de la forma más sencilla posible.

Como lenguaje de programación para el procesamiento de los datos se ha utilizado Python y librerías propias de un entorno Big Data (PySpark, por ejemplo). El análisis previo de los datos se ha realizado en R.



Esquema de la captación de datos

4. Resultados

El resultado final del proyecto es una herramienta de ingesta de datos multifuente para nutrir de información un sistema de predicción de rutas y visualización de información relacionada con las rutas realizadas con bicicletas.

El enfoque a la seguridad del usuario en este proyecto determina la necesidad de adaptar datos relativos a accidentes, tráfico y polución, tanto en tiempo real como en diferido. El tratamiento de estos datos en esta herramienta elaborada posibilita la utilización de esta información.

5. Conclusiones

La ingesta de datos realizada ha sido manual, siendo la automatización de la misma un posible objetivo futuro.

Por otro lado, las técnicas de procesamiento del Raw Data han tenido que realizarse de una forma artesanal, sobre todo debido a la naturaleza de los datos del ayuntamiento de Madrid, tanto porque la estructura de los datos era poco beneficiosa para su procesamiento en un entorno Big Data, como porque estos datos traían consigo una gran cantidad de errores.

DATA ACQUISITION FOR A BIG DATA PROJECT

Author: Rodríguez-Mon Barrera, Beltrán.
Supervisor: Contreras Bárcena, David.
Collaborating Entity: ICAI – Universidad Pontificia Comillas

ABSTRACT

Keywords: API, WebScrapping, Big Data, Raw Data

1. Introduction

Nowadays, a lot of data is generated by a large number of elements used daily by people. This information can be processed in order to make decisions according to the final objectives of a project.

It is of great importance to provide a Big Data system with a diversity of data that complements each other to nourish a decision system of sufficient tools so that it is able to get to best possible solution.

2. Project definition

The project's objective is the elaboration of a Data Acquisition System for a Big Data structure. The data will be ingested using technologies such as Web Scrapping, access to real-time APIs or other types of multi-source data ingestion.

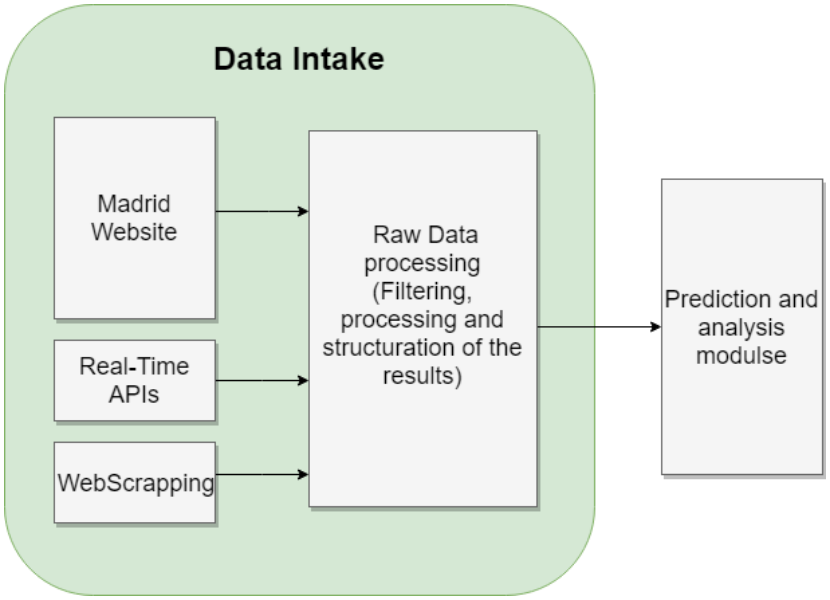
The main source of data will be the open repository provided by the city of Madrid website, focusing the data to complement a set of journeys collected during several months of the years 2018 and 2019.

3. Model/System/Tools description

The data collection will be done manually from various sources (the city of Madrid, principally, but data from other sources will also be adapted).

Then, a data cleansing will be carried out in a Big Data Cluster offered by the ICAI University, reducing the data to only the necessary information for later stages of the Project, and adapting them so that their subsequent process is as simple as possible.

Python and some libraries related to the Big Data environment (PySpark, for instance) have been used during this project. Also, the initial data analysis has been carried out in R.



Data intake schema

4. Results

The final result of the project is a data intake tool for a route prediction system and the display of information related to bicycle travels.

The approach to the user safety in this project determines the need to adapt the data related to accidents, traffic and pollution, both in real time and in deferred time. The treatment given in this tool allows the use of this information.

5. Conclusions

The data intake has been manual, being the automation of the same a possible future objective.

On the other hand, the Raw Data processing techniques have been carried out in an artisanal way, mainly due to the nature of the Madrid website data, both because the data structure was not very beneficial for a Big Data processing, as because these data brought with them a lot of errors.

Índice de la memoria

Índice de ilustraciones

Índice de tablas

Capítulo 1. INTRODUCCIÓN

La informatización de una gran cantidad de elementos utilizados en el día a día de las personas está provocando que se generen multitud de datos. Esta información puede ser procesada para acabar generando diversas tomas de decisiones según el objetivo final que se tome

Es de una gran importancia proveer un sistema de una diversidad de datos que se complementen entre sí para nutrir a un sistema de decisión de las herramientas suficientes para que este pueda tomar la mejor resolución posible.

Este proyecto trata de la elaboración de un sistema de ingesta de datos para su posterior integración y procesamiento en una infraestructura Big Data. La ingesta de datos se realizará mediante tecnologías de captación de datos tales como el WebScrapping, el acceso a APIs en tiempo real u otros tipos de ingesta de datos multifuente.

La fuente principal de datos será el repositorio abierto de datos del ayuntamiento de Madrid. El foco central de la captación de datos serán los archivos de trayectos de bicicletas de la compañía pública “Bicimad”, en la que se nos ofrecen datos tales como información GPS de los viajes, la duración de estos y las estaciones de origen y destino, entre otros.

Estos datos se complementarán con otras fuentes, ya sean también ofrecidas por el ayuntamiento de Madrid (como información del tráfico, o informes de accidentes) o por otros proveedores (como bases de datos de uso abierto del tiempo meteorológico, por ejemplo).

La intención de este proceso de ingesta de datos es proporcionar una muestra diversa de datos para poder aplicar sobre los mismos algoritmos de aprendizaje y sacar modelos que ofrezcan soluciones al usuario.

El enfoque de esta aplicación hacia la seguridad supone un cambio en la visión tradicional que se tiene en el desarrollo de aplicaciones de trazado de rutas, que principalmente se centran en la minimización del tiempo del trayecto. Tras el análisis previo de los accidentes relacionados con bicicletas en Madrid, se vio que estos accidentes se concentran en calles en concreto, y no se distribuyen de forma equitativa. De este modo, considerando todos los parámetros que pueden definir un accidente (hora, día, tiempo atmosférico, etc.), y añadiendo los datos en tiempo real comparables con los anteriores, se podría establecer el camino óptimo para un usuario minimizara el tiempo de trayecto a la vez que maximiza la seguridad de este.

Ilustración 1. Mapa de calor de accidentes de Madrid 2018

Como se aprecia en la ilustración 1, los accidentes se concentran en calles en concreto, y teniendo en cuenta información sobre la fecha y hora, se puede aconsejar al usuario evitar ciertas calles donde la concentración de accidentes es más frecuente.

Para la preparación de los datos para el procesamiento de estos accederá a repositorios de datos de diversas fuentes y tipos, se recolectarán y filtrarán para mantener únicamente aquella información que sea necesaria para el procesamiento, y se almacenarán en un clúster para el posterior acceso. Además, se complementarán los datos con otras fuentes para nutrir debidamente el sistema.

Capítulo 2. DESCRIPCIÓN DE LAS TECNOLOGÍAS

Big Data:

Colección de datos grande y compleja, difícil de manejar con mecanismos tradicionales. Se puede resumir las características del Big Data mediante las "5 Vs" (Volumen, Velocidad, Variedad, Veracidad y Valor).

API (Application Programming Interface):

Comunicación entre aplicaciones para interactuar entre sí. En el contexto de este proyecto se entiende como el establecimiento de una comunicación para la adquisición de datos y funciones de uso extendido.

WebScrapping:

Extracción de información mediante software de una página web. Se puede acceder a una página web simulando un "acceso humano" y analizar la información que ésta te proporciona mediante, por ejemplo, la navegación en un HTML por sus etiquetas.

Raw Data:

Dato recogido directamente de la fuente que lo genera y que no ha sido transformado de ninguna forma.

Clúster:

Conjunto de máquinas que utilizan un hardware compartido y se comportan como si fueran un único ordenador. Tienen servicios compartidos y están monitorizados entre sí.

Hdfs (Hadoop Distributed File System):

Sistema de almacenamiento de ficheros de manera distribuida. Está caracterizado por tener una topología de Maestro-Esclavo y por tener un almacenamiento redundante, dividiendo los archivos en bloques y almacenándolos en distintas localidades físicas, proporcionando así tolerancia a fallos.

Está compuesto por un Namenode, cuyo objetivo es administrar los Esclavos y conocer la localización de los datos, y Datanodes, que almacenan datos y realizan solicitudes de lectura y escritura.

Yarn (Yet Another Resource Negotiator):

Plataforma de negociación de recursos de un clúster Hadoop. Yarn se encarga de distribuir los procesos por el clúster Hadoop hacia los nodos en los que se encuentran los datos necesarios, recoger el resultado obtenido y notificarlo a el programa que lo solicitó.

Está compuesto por ResourceManager y NodeManagers; el primero recibe una instrucción con el proceso a realizar y se lo distribuye a los segundos, para que al finalizar devuelvan los resultados obtenidos.

PySpark:

Librería Python que adapta Apache Spark, framework de computación open-source en clústeres. Ofrece una lectura paralelizada de datos distribuidos trabajando en un entorno tolerante a fallos.

Spark Dataframe:

Organización de datos distribuidos en columnas, equivalente en concepto a una tabla en bases de datos relacionales, pero con un acceso optimizado.

BeautifulSoup:

Librería Python para acceder a datos provenientes de HTML o XML, mediante la indexación por etiquetas.

Jupyter Notebook:

Aplicación web para la creación y compartición de código de análisis, visualización, creación de modelos…

Capítulo 3. ESTADO DE LA CUESTIÓN

En la actualidad existen multitud de aplicaciones, tanto en formato web como sólidas, para el establecimiento de rutas de viaje desde un punto de origen a un destino en concreto para optimizar la duración de este. Estas aplicaciones utilizan distintas tecnologías, tales como GPS o señalización mediante protocolos propios de Internet, más el uso de algoritmos para determinar cómo minimizar el tiempo utilizando las calzadas disponibles.

A su vez, hay una gran cantidad de datos de acceso abierto ofrecidos por organizaciones, tanto de históricos recopilados para el análisis posterior como de datos en tiempo real sobre la situación en un momento exacto, ya sea de elementos en concreto o datos generales como el tiempo atmosférico. Otro tipo de datos que acaban siendo disponibles de forma abierta son aquellos generados por propios usuarios de Internet (sin tener que ser participantes de una organización, pero sí de algún grupo o red social).

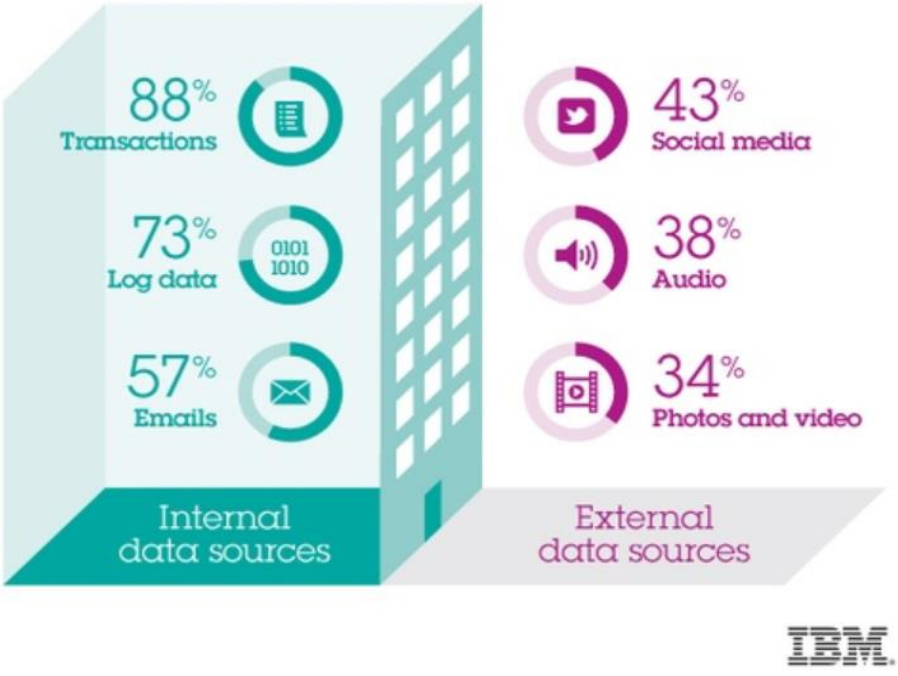


Ilustración 2. Origen de los datos [1]

3.1 LOS DATOS EN BIG DATA

El formato de estos datos varía mucho dependiendo tanto de su naturaleza como del objetivo por el cual han sido creados. Por ejemplo, en una empresa de finanzas es probable que la mayoría de los datos estén guardados en archivos generados por software utilizado en ese campo (Excel, por ejemplo). Por otro lado, una empresa enfocada en la informática, en comunicaciones o en servicios web utilizará otro tipo de archivos y formatos para guardar toda la información relativa a su área de trabajo (.json, …).

Esto implica que la adquisición de datos se haya de realizar de una forma concreta teniendo en cuenta tanto el origen de la información como el destino u objetivo que se les quiera dar. Además, su tratamiento dependerá tanto de los factores mencionados anteriormente como de otros; por ejemplo, el tamaño de los datos implicará la obligación de tener que imponer un tipo arquitectura Big Data y una lógica de procesamiento sobre otra.

También cabe destacar la importancia de la estructuración de un dato, pues éstos no se procesarán de la misma forma. Por un lado, tenemos los datos estructurados, es decir, aquellos cuyo formato es fijo. Un ejemplo claro de estos tipos de datos serían los archivos .xls (Excel) con formato de tabla, que se rellenan con tuplas para guardar la información. Por otro lado, tenemos los datos semiestructurados, que, aunque su estructura no esté formalizada, poseen cierta organización mediante etiquetas o marcadores (ejemplos de este tipo de dato serían los .json o .xml). Por último, están los datos no estructurados, que para su procesamiento se tendrán que convertir en un dato de alguno de los otros dos tipos mencionados anteriormente mediante el conocimiento de la creación del dato. Poniendo otro ejemplo, un archivo .txt en el que la información se guarda en líneas no tiene estructura en el propio archivo, pero mediante el conocimiento de la creación de estas líneas y su origen, podríamos transformarlos para su posterior análisis.

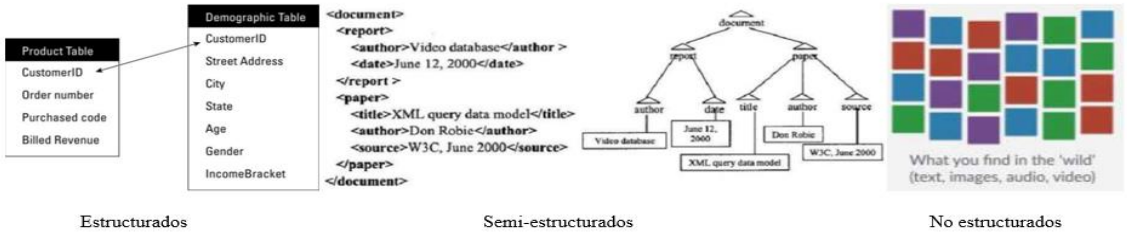


Ilustración 3. Representación de los tipos de datos [2]

La importancia de la calidad del dato es vital, puesto que el objetivo principal es conseguir crear conocimiento a partir de un conjunto de datos puros (sin tratar, y puede que sin una relación directa y clara con el objetivo que se busca). Tanto para lo bueno como para lo malo, desde que un dato puro entra en el sistema hasta que se presenta un resultado generado por un modelo, existe la necesidad de que el conocimiento sobre el dato adquirido por cada integrante, debido al trabajo que hay detrás de la adquisición, de un filtrado o de un procesamiento que implique cambiar la naturaleza del mismo, se transmita de la forma idónea, consiguiendo así crear una comunicación eficaz en todo el proceso Big Data que maximice las capacidades de trabajo en el procesamiento de una o varias fuentes de datos [3].

3.2 ARQUITECTURAS DE ALMACENAMIENTO DE DATOS EN BIG DATA

Por otro lado, en las estructuras de almacenamiento de datos en Big Data de hoy en día se utilizan diferentes arquitecturas, pero todas ellas buscando maximizar características como el acceso en alta velocidad a los datos, o el análisis en tiempo real de los mismos. El almacenaje de datos es un elemento presente en una gran cantidad de empresas hoy en día, que ofrecen sus datos a usuarios finales a través de una interfaz para que estos puedan analizarlos. Estos almacenes han ido evolucionando desde sus primeras implementaciones en el último tercio del siglo XX hasta la actualidad.

En 1988 IBM Systems Journal publicó “An architecture for a Business Information System” [4], definiendo la problemática y futura solución que tendría el almacenamiento de

datos de las empresas. Además, la llegada de los ordenadores personales a las empresas, y la utilización de "Software de oficina" (Word, Excel…) permitió un almacenaje masivo e información de forma sencilla. A partir de ahí, esta acumulación de datos se puso a disposición de desarrolladores para su uso en toma de decisiones.

Poniendo un ejemplo actual, el ayuntamiento de Madrid ofrece de forma abierta a cualquier usuario sets de datos sobre distinta índole para que puedan set analizados. La estructura de los datos, como se ha ido mencionando anteriormente, es muy diversa, sobre todo porque en la actualidad los datos generados por una gran cantidad de organizaciones no están creados con la intención de que sean analizados por otro grupo de trabajo, sino que son datos del ámbito en el que se mueven que deciden abrir públicamente para quien los desee.

En cuanto al trabajo con estos datos, existen multitud de estructuras y arquitecturas sobre las que crear sistemas de almacenamiento y procesamiento masivo de datos. Uno de estos Frameworks es Hadoop, sistema de datos distribuido licenciado bajo licencia Apache v2, desarrollado basándose en el artículo de Google sobre GFS[5] y el sistema MapReduce[6].

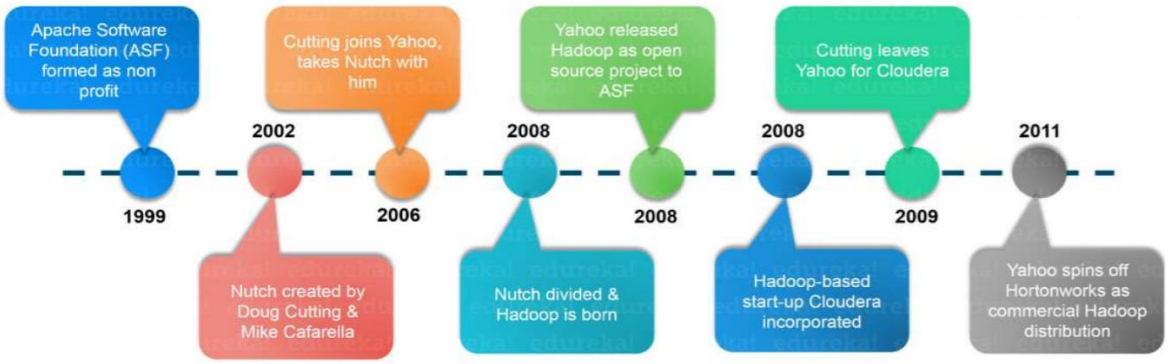


Ilustración 4. Evolución de Hadoop [2]

Así como Hadoop, existen distintas arquitecturas de Big Data utilizadas para la ingesta y análisis de datos. La elección de una sobre otra viene dada tanto por las características del dato, el tipo de análisis deseado, u otros factores externos como el presupuesto disponible. Como ejemplo, no tendría sentido utilizar una arquitectura de IoT

(Internet of Things) para el análisis de históricos de gran volumen. Aun así, los componentes de las distintas arquitecturas de dato no difieren en gran medida entre ellas, pero sí lo hace el modo en el que estos componentes se utilizan. En la siguiente figura (Ilustración 5) se muestran las diferencias entre una arquitectura genérica y una IoT, viendo cómo se comparten elementos en ambas, pero que su utilización/posición difiere.

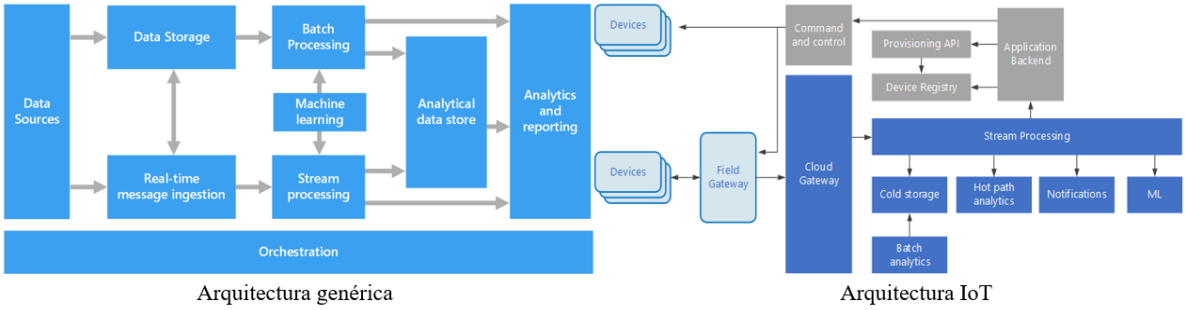


Ilustración 5. Arquitecturas Big Data [7]

3.3 APLICACIONES DE TRAZADO DE RUTAS

El uso de estos datos está muy extendido en multitud de aplicaciones. Desde las propias empresas, que toman decisiones utilizando su propia generación de datos, hasta aplicaciones que utilizan estos datos para la generación de modelos. Un claro ejemplo de esto sería Google Maps, que tanto se nutre de información que ha generado por sí misma, como las carreteras o posiciones GPS de cientos de miles de establecimientos, para establecer cómo llegar de un origen a un destino, y ofrecer alternativas según otras fuentes le proporcionen información adicional para tomar una mejor decisión, como el tiempo atmosférico en el momento de hacer la consulta, o información sobre tráfico y accidentes.

Google Maps es una herramienta utilizada en el día a día por una gran cantidad de usuarios, con una aplicación tanto de estado sólido como web que te da la posibilidad de calcular una ruta de viaje. El cálculo de rutas que se lleva a cabo en esta aplicación se basa en una representación del mapa en forma de grafos dirigidos, siendo cada uno de los cruces un vértice y las calles las aristas. Tras esto se utilizan algoritmos basados en el Algoritmo de Dijkstra para que, mediante el cálculo del camino más corto considerando grafos con pesos,

para acabar representando la ruta que se nos ha mostrado [8][9]. Los problemas añadidos con los que una aplicación tiene que lidiar es, aunque la teoría marca cómo se puede calcular la ruta, existen factores externos que provocan cambios de forma dinámica. Si hay un atasco, o una ruta está cortada de forma temporal, Google deberá ser capaz de captar estos nuevos datos mediante una ingesta multifuente y procesarlos para volver a ponderar los pesos de las rutas y recalcular cuál es la más optima.

Un posible algoritmo utilizado para el cálculo de rutas es A* (A star), que se sirve de algoritmos heurísticos (que nunca sobreestiman un coste de una arista) para calcular la expansión de un camino. A* es una modificación de Dijkstra optimizado para llegar a un destino concreto, basado en analizar primero aquellos posibles caminos que sean más prometedores.

Otra herramienta utilizada en este ámbito a tratar es la aplicación oficial de BiciMAD, que ofrece al usuario forma de pago de su sistema de alquiler de bicis, información de sus estaciones y creación de históricos, entre otras cosas. La ventaja que tiene la aplicación del sistema de alquiler público es el acceso a los datos de primera mano, como por ejemplo información de todas las estaciones de bicicletas, gracias a la infraestructura pública que utilizan. Aun así, esta aplicación no cuenta con un método de cálculo de rutas (o integración de uno), por lo que por sí sola no serviría para el propósito marcado.

Madbike supone una integración de estas dos aplicaciones mencionadas anteriormente, utilizando la API de la aplicación de BiciMAD e integrando en esta app la aplicación de Google Maps. La aplicación creada por Álex Rupérez y Javier Muñoz une todos los datos abiertos que ofrece el ayuntamiento de Madrid desde su web y una forma de calcular rutas en una misma interfaz, haciendo el acceso a la información que hay tras la gran cantidad de datos generados por el ayuntamiento más sencillo para el usuario final.

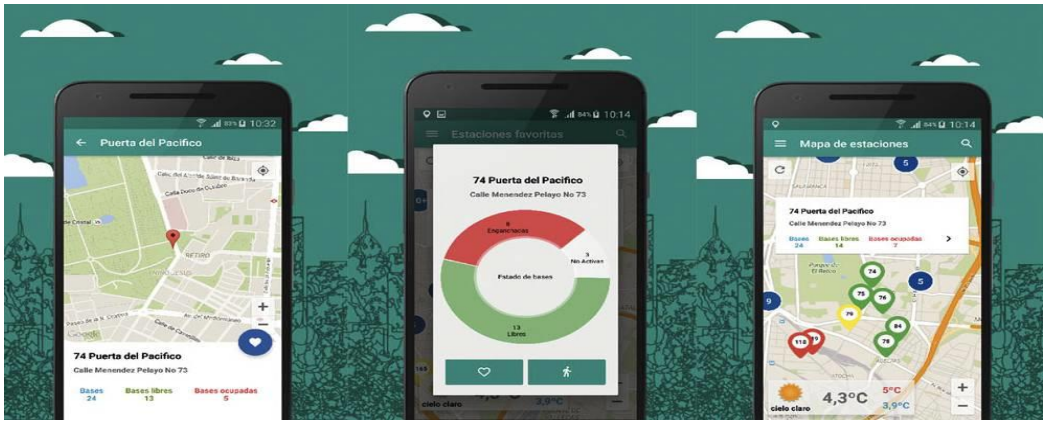


Ilustración 6. Interfaz de Madbike [10]

3.4 GENERACIÓN DE NUEVA INFORMACIÓN EN INTERNET

A su vez, estamos en una situación en la que la información sobre diversos eventos fluye rápidamente a través de Internet. El acceso a periódicos digitales, y el uso masivo de redes sociales proporcionan una gran cantidad de información a la que cualquier usuario puede acceder. En estas dos fuentes de datos hay que tener en cuenta dos características fundamentales, que es posible que se contrapongan: La rapidez con la que se propaga una información con la veracidad de esta, mencionado en [11].

Debido a esto los sistemas han de ser capaces de comprobar, de una forma rápida, que una noticia generada en momentos después de que el evento haya ocurrido, es veraz, para poder tomar decisiones lo más rápido posible.

Capítulo 4. DEFINICIÓN DEL TRABAJO

4.1 DATOS UTILIZADOS Y JUSTIFICACIÓN

El objetivo del proyecto, como se mencionó anteriormente, se puede resumir en generar rutas para el usuario que, aparte de considerar la distancia entre origen y destino, también tenga en cuenta la seguridad. Es por esto que en este proyecto se han buscado integrar una diversidad de datos que nos puedan ayudar a conseguir estos objetivos y, considerando las fuentes disponibles (el repositorio abierto del ayuntamiento, APIs de terceros de libre uso e información proporcionada directamente en páginas web) se ha llegado a la integración de los datos que serán descritos a continuación.

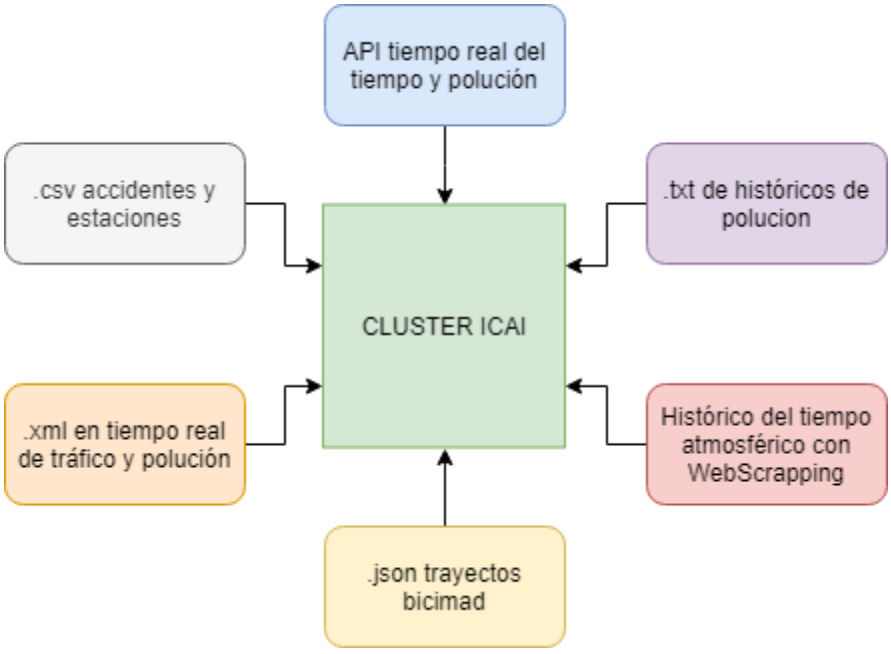


Ilustración 7. Datos recogidos para el proyecto.

4.1.1 TRAYECTOS DE BICIMAD

El core de todo este proyecto Big Data realizado se mueve alrededor de este set de datos que nos proporciona el ayuntamiento de Madrid sobre las bicicletas de BiciMAD, dándonos información sobre el tiempo de uso, tipo de usuario que utiliza el servicio… pero sobre todo del trayecto que ha realizado.

Los más de 3 millones y medio de trayectos recogidos en estos archivos de datos son vitales para entrenar adecuadamente el sistema utilizado para el cálculo de rutas. Además de eso, el resto de información que nos proporciona este set de datos puede ser utilizado para realizar estadísticas sobre el uso de las bicicletas de la compañía pública en Madrid, tanto para la propia organización como para los usuarios (conocer la utilización de cada estación, por ejemplo).

4.1.2 HISTÓRICO DE ACCIDENTES

El objetivo principal de este proyecto es poner en relación los trayectos realizados en bicicleta con la seguridad vial, por lo que era necesario nutrir al sistema de información de los accidentes relacionados con bicicletas.

Aunque se podría presuponer, era necesario comprobar la relación entre los accidentes y periodos de la semana y año, o también ver si se concentraban en carreteras o elementos viales en concreto, para poder sacar puntos peligrosos de tráfico que nuestro sistema ha de intentar evitar.

Por otro lado, con la información que estos históricos nos proporcionan podemos conocer la situación atmosférica que se daba cuando ocurrieron los accidentes, y cruzando esta información con la del tiempo atmosférico, poder sacar conclusiones sobre su relación.

Debido a esto se estableció que era importante agregar estos datos a nuestro sistema y transformarlos (como se explicará más adelante) para que se adapten adecuadamente.

4.1.3 ESTACIONES

Dado que todo viaje recogido en los datos de trayectos de BiciMAD tienen como origen y destino una estación, pero que dentro de la información de trayecto no está recogida como coordenadas del mapa, sino como un ID que identifica la estación, se tuvo que integrar en el sistema la información de cada estación para construir completamente la ruta realizada.

4.1.4 HISTÓRICO DE TIEMPO ATMOSFÉRICO

La clara implicación del tiempo atmosférico con la posibilidad de que exista un accidente crea la necesidad de construir un histórico del tiempo con la información disponible del tiempo en Internet.

El principal problema existente era la escasez de información de históricos del tiempo que estuvieran disponibles de forma gratuita. Es por esto por lo que se decidió crear un histórico a partir de herramientas de WebScrapping y de la página web ogimet.com.

Esta página web ofrece de forma gratuita esta información estructurada en días y dividida según estaciones meteorológicas distribuidas por todo el mundo. La información que ofrecen está generada a partir de información de la NOAA (National Oceanic and Atmospheric Administration) y de software libre para su tratamiento.

4.1.5 HISTÓRICO DE POLUCIÓN

La contaminación es un factor que cada día está adquiriendo una gran importancia en nuestra sociedad, y en este proyecto se ha querido ver la relación que podría existir entre los trayectos de bicicletas y la acumulación de polución en distintas áreas de la ciudad, por lo que se ha decidido integrar en el sistema los históricos recogidos en formato .txt en el repositorio del ayuntamiento.

Estos datos proporcionan una gran cantidad de información sobre polución en Madrid tanto de las numerosas estaciones de toma de medidas distribuidas por todo Madrid como por las distintas medidas que éstas toman.

4.1.6 INFORMACIÓN EN TIEMPO REAL DEL TIEMPO ATMOSFÉRICO

En una aplicación como la que se busca desarrollar en nuestro proyecto es crucial tener información en tiempo real de diversos factores que afecten al comportamiento del sistema.

El uso de una API web como recurso para conseguir información del tiempo se debe a que en la adquisición de esta información se busca maximizar su velocidad, y se decide la implementación en el sistema de la API de uso abierto de OpenWeatherMap para conseguir información del tiempo en Madrid. La información adquirida son datos correspondientes a un único punto de medida, pero debido a que el área de trabajo de la aplicación corresponde a Madrid centro, es suficiente.

4.1.7 INFORMACIÓN EN TIEMPO REAL DEL TRÁFICO

Obtener información en tiempo real del tráfico en las calzadas ayudaría al sistema a tomar decisiones para evitar posibles zonas de peligro y zonas de atasco. Con información de intensidad de tráfico y el ritmo que llevan los vehículos por determinadas zonas, se podría obtener las mejores rutas para el momento en el que se solicitan.

Se decide utilizar la información oficial del ayuntamiento respecto al tráfico dado en formato web debido tanto a su rápida actualización (entre 5 y 10 minutos) como a la fiabilidad que proporciona el conocimiento de la fuente de origen del dato.

4.1.8 INFORMACIÓN EN TIEMPO REAL DE POLUCIÓN

Como se explicó anteriormente, se deseaba integrar en este sistema la información de polución para conocer la relación con los datos. A la hora de adquirir datos sobre la polución, y debido a la diversidad de fuentes existentes, se decidió integrar dos fuentes distintas de estos datos: una de ellas es un archivo de acceso web del ayuntamiento, con información en tiempo real de todas las estaciones funcionando en este momento, y la otra una API en tiempo real proporcionada por aqicn.org, que proporciona datos sobre polución

en una de sus estaciones en Madrid. El uso de una opción sobe otra dependerá de la necesidad de anteponer el tiempo de adquisición de la información sobre la cantidad de esta.

4.2 OBJETIVOS

El objetivo de este proyecto es la ingesta multifuente de datos para su integración en un sistema Big Data. La aplicación a diseñar integrará los datos previamente recolectados, filtrados y cargados al sistema para, mediante un algoritmo basado en Machine Learning, tomar decisiones según las entradas de usuario.

De esta forma, los objetivos de este proyecto quedan marcados de la siguiente forma:

PRIMARIOS:

- Familiarizarse con la fuente principal de datos sobre la que se basa el proyecto: Los datos de trayectos de Bicimad y los ficheros de accidentes disponibles de la comunidad de Madrid. Realizar un análisis para sacar conclusiones sobre qué tipo de fuentes de datos son las más convenientes para nutrir nuestro sistema

- Integración de históricos de datos que se utilizarán para el aprendizaje del modelo. Captación de distintos tipos de datos provenientes de fuentes diversas.

- Uso de APIs en tiempo real para utilizar como parámetros de entrada del modelo, como información del tiempo atmosférico o el tráfico.

SECUNDARIOS:

- Análisis de la API de Twitter para conseguir datos dinámicos para conseguir detectar eventos aún no notificados en las versiones digitales de los medios, o que usuarios publiquen.

4.3 METODOLOGÍA

Los primeros meses del proyecto se utilizarán para familiarizarse con el trabajo en un entorno Big Data. Primero de todo se realiza un análisis estadístico en pequeña escala de las variables principales (el histórico de los trayectos de bicicletas de Bicimad) utilizando R. Este análisis se llevará posteriormente a gran escala integrando todos los datos en el clúster y, mediante funciones de Python, se sacarán conclusiones sobre el conjunto total de los datos.

Tras esto se agregarán otras fuentes de datos y se filtrarán utilizando PySpark según las necesidades correspondientes para integrar en la infraestructura Big Data únicamente aquella información que vaya a ser de utilidad en los siguientes pasos del proyecto.

Posteriormente se utilizarán funciones de Python de captación de datos mediante WebScrapping, con el objetivo de crear una estructura PySpark homogénea para que el modelo posterior pueda analizar y tomar decisiones.

Por último, se accederán a APIs en tiempo real para poder apoyar el modelo en datos altamente relacionados con la funcionalidad de la aplicación (como el tráfico o el tiempo atmosférico, por ejemplo).

4.4 PLANIFICACIÓN Y ESTIMACIÓN ECONÓMICA

4.4.1 PLANIFICACIÓN

Tabla 1. Planificación del proyecto

4.4.2 ESTIMACIÓN ECONÓMICA

El presupuesto necesario para la ejecución del proyecto Big Data, considerando tanto ingesta, procesamiento y análisis y visualización de datos, se puede dividir en dos partes claramente diferenciadas: la mano de obra y los equipos para el procesamiento informático.

Elementos utilizados	**€/ud**	**uds**	**total**
Clúster ICAI	80.000	1	80.000
Xiaomi mi Air 13,3"	899	1	899
Dell Inspiron 13 7000 series	761	1	761
Lenovo Y50-70	965	1	965
		TOTAL	82.625

Partes del proyecto	**horas**	**€/hora**	**total**
Análisis de los trayectos	100	7*3	2100
Programación de la Ingesta de datos multifuente	300	8,25	2475
Predicción de rutas	350	9	3150
Visualización de datos	300	8,5	2550
		TOTAL	10275

En la etapa de análisis de los trayectos de BiciMAD se consideran los sueldos de los tres integrantes del proyecto, quedando representado en la tabla de estimación del coste de la mano de obra como el sueldo por hora de un integrante multiplicado por el número de integrantes (7*3).

Capítulo 5. DESARROLLO DEL PROYECTO

Como se ha explicado anteriormente, el proyecto se ha basado en la recopilación de datos que se pueden dividir según su contenido y su formato.

Antes de entrar a explicar cada tipo de dato recogido para integrar en el sistema, cabe destacar la importancia de la utilización de un clúster Big Data en este tipo de proyecto: El hecho de trabajar con una metodología Big Data no implica únicamente la existencia de una gran cantidad de datos, sino que hay que alejarse de los análisis convencionales (como el primero que se realizó a pequeña escala en este proyecto, que se explicará más adelante) para enfocarse en uno a gran escala que implican el uso de tecnologías propias de esta área.

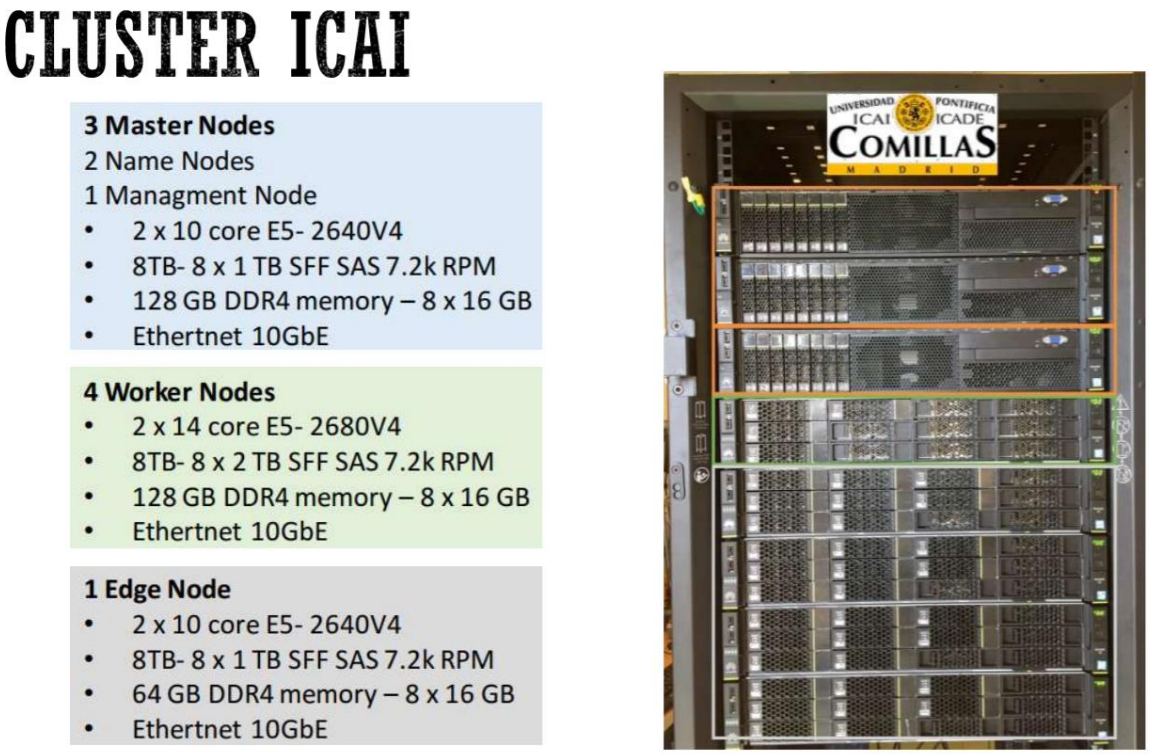


Ilustración 8. Clúster de ICAI

Es de vital importancia utilizar lo máximo posible las estructuras y funciones propias de un sistema distribuido. Dado que el lenguaje de programación utilizado en este proyecto

es Python se hará uso de la librería PySpark, y de sus estructuras (principalmente los dataframes) y funciones. Spark es un marco de desarrollo para entornos Big Data que está montado por encima de Hadoop, que utilizamos para tratar los datos.

Ilustración 9. Spark

Como es propio de Hadoop, los datos se mantienen estáticos y es el proceso el que se mueve a ellos, por lo que hay que procurar en lo máximo de lo posible mantener los datos en el sistema HDFS y no moverlos a la memoria local, para poder utilizar los recursos que el clúster nos proporciona.

5.1 ANÁLISIS DE LOS TRAYECTOS DE BICIMAD

El análisis de los trayectos de bicicleta de la compañía BiciMAD constituye la primera etapa realizada en este proyecto Big Data. El análisis fue realizado por todo el grupo y los principales objetivos de éste fue tanto la familiarización con las variables de este bloque de datos como la búsqueda en conjunto de posibles retos que se podían afrontar en un proyecto de este estilo utilizando como datos principales estos trayectos de bicicletas.

```
root
 |-- _id: struct (nullable = true)
 |    |-- $oid: string (nullable = true)
 |-- ageRange: long (nullable = true)
 |-- idplug_base: long (nullable = true)
 |-- idplug_station: long (nullable = true)
 |-- idunplug_base: long (nullable = true)
 |-- idunplug_station: long (nullable = true)
 |-- track: struct (nullable = true)
 |    |-- features: array (nullable = true)
 |    |    |-- element: struct (containsNull = true)
 |    |    |    |-- geometry: struct (nullable = true)
 |    |    |    |    |-- coordinates: array (nullable = true)
 |    |    |    |    |    |-- element: double (containsNull = true)
 |    |    |    |    |-- type: string (nullable = true)
 |    |    |    |-- properties: struct (nullable = true)
 |    |    |    |    |-- secondsfromstart: long (nullable = true)
 |    |    |    |    |-- speed: double (nullable = true)
 |    |    |    |    |-- var: string (nullable = true)
 |    |    |    |-- type: string (nullable = true)
 |    |-- type: string (nullable = true)
 |-- travel_time: long (nullable = true)
 |-- unplug_hourTime: struct (nullable = true)
 |    |-- $date: string (nullable = true)
 |-- user_day_code: string (nullable = true)
 |-- user_type: long (nullable = true)
 |-- zip_code: string (nullable = true)
```

Ilustración 10. Información de las variables de los trayectos.

Primero de todo cabe explicar cada una de las variables de este set de datos dado que, como se ha mencionado, es el set principal sobre el que se enfoca el proyecto.

5.1.1 EXPLICACIÓN DE LOS DATOS

5.1.1.1 ID

Identificación de un trayecto concreto.

Este valor no se repite durante toda la muestra de datos, lo que hace presuponer que se realizan mediante un hash de variables para generar un viaje.

5.1.1.2 USERDAYCODE

Código generado para un usuario en un día concreto.

El objetivo de esta variable es realizar estudios estadísticos de tendencias diarias de un usuario, por lo que se repite cada día el número de veces que un mismo usuario haya realizado un trayecto.

5.1.1.3 PLUGBASE

Identificador de la base en la que se guarda la bicicleta.

La única interpretación posible de esta variable sería para conocer si hay algún pivote para guardar la bicicleta que haga mal contacto.

5.1.1.4 UNPLUGBASE

Identificador de la base de la que se saca la bicicleta.

Igual que “plugbase”, este dato tiene poco valor.

5.1.1.5 TRACK

Detalle del recorrido realizado por la bicicleta entre la estación origen y destino. Esta variable contiene los datos de las coordenadas (longitud, latitud), dirección de la posición,

velocidad en ese momento, tiempo en el que se recoge la muestra (respecto al inicio -> t=0). Estos datos son recogidos en un intervalo regular de 60 segundos.

Esta variable es el dato principal utilizado para la creación del modelo de predicción posterior. El principal problema de esta variable es que es la única que puede tener valores nulos sin que sea por causas excepcionales. Durante los meses de noviembre de 2017 y abril de 2018 los datos directamente no poseen la variable 'TRACK', lo que supone una pérdida importante de información para el modelo, y en el resto de meses

De un total de 5.333.647 trayectos, únicamente 1.878.959 tienen información de las coordenadas por las que ha sucedido, lo que significa una pérdida de 64,77% sobre esta información, y descartando aquellos meses en los que no existe campo de TRACK, únicamente un 50,51% de los datos tiene un valor distinto de nulo (sobre un total de 3.719.665 datos). Aun así, tener cerca de 2 millones de datos con esta información del trayecto son suficientes para generar un modelo eficaz.

5.1.1.6 USERTYPE

Identificación del tipo de usuario que ha realizado un movimiento en concreto. Estos valores son:

0: Usuario no identificado.

1: Usuario poseedor de un bono anual.

2: Usuario ocasional.

3: Trabajador de mantenimiento.

Existe la posibilidad de dividir entre mantenimiento y el resto para obtener información sobre aquellos viajes realizados por los usuarios de BiciMAD y, relacionándolo con otras variables, se pueden sacar relaciones sobre el tipo de recorridos de un determinado sector de usuarios.

5.1.1.7 TRAVELTIME

Tiempo total, en segundo, entre el desenganche y el enganche de la bicicleta.

Con la información de esta variable se puede identificar que algunos viajes se realizan en intervalos de tiempo anómalos, desde tiempos de pocos segundos y tiempos muy altos (de la escala de varios días), que posiblemente se corresponda tanto a bicicletas que se han desenganchado de su base y vuelto a enganchar (ya sea de forma accidental o intencionada) como a bicicletas que se han perdido o dejado abandonadas, tratando éstos por separados para otro tipo de análisis.

Existe la posibilidad de establecer outliers para filtrar ciertos trayectos que se tomen como imposibles o que no sirvan para el entrenamiento del modelo posterior.

5.1.1.8 PLUGSTATION

Identificador de la estación de destino.

Se pueden ver las estaciones más utilizadas como destino, asociarla a una región de Madrid, y sacar visualizaciones sobre esto.

5.1.1.9 UNPLGUSTATION

Identificador de la estación de origen.

Vista de las estaciones en las que se sacan más bicicletas. Como con “plugstation”, se puede analizar la región en la que más bicis se necesitan.

5.1.1.10 AGERANGE

Identificación del rango de edad del usuario que ha realizado el movimiento:

0: Rango de edad no identificado.

1: Usuario entre 0 y 16 años.

2: Usuario entre 17 y 18 años.

3: Usuario entre 19 y 26 años.

4: Usuario entre 27 y 40 años.

5: Usuario entre 41 y 65 años.

6: Usuario de más de 66 años.

5.1.1.11 UNPLUGHOURTIME

Franja horaria en la que se ha realizado el trayecto. El valor almacenado es la fecha AAAA-MM-DD + la hora, pero no se da información de minutos y segundos por cuestiones de anonimato. Debido a esto, todos los trayectos de un día que se inicien en la misma hora tendrán el mismo valor en esta variable.

Con esta información se pueden analizar los días del mes de más afluencia (fines de semana / entre semana), hora del día que más bicis se requieren por los usuarios… Este dato también se puede unir con el de tipo de usuario para sacar qué utilidad le da la gente a las bicis.

5.1.1.12 ZIPCODE

Código postal del usuario que ha realizado el trayecto.

Se podría analizar la región dentro de Madrid en el que hay un amplio número de usuarios, comparando a su vez con el número de estaciones disponibles, para ver si es necesario instalar más estaciones, o si en una zona no se dan uso.

5.1.2 ANÁLISIS A PEQUEÑA ESCALA

Como se ha ido explicando a lo largo de ese apartado, existen muchas posibilidades de enfoque del proyecto según la información que nos proporcionan estas variables, por lo que el primer objetivo fue la realización de un análisis a pequeña escala de las variables de

los trayectos de bicicletas. Este análisis se realizó utilizando el lenguaje de R, software abierto para computación estadística.

Para este análisis se utilizó el set de datos correspondiente a abril de 2017 (elegido de forma aleatoria) siendo el primer objetivo de este análisis la obtención datos estadísticos sobre las variables para tomar decisiones y conocer la naturaleza de las variables.

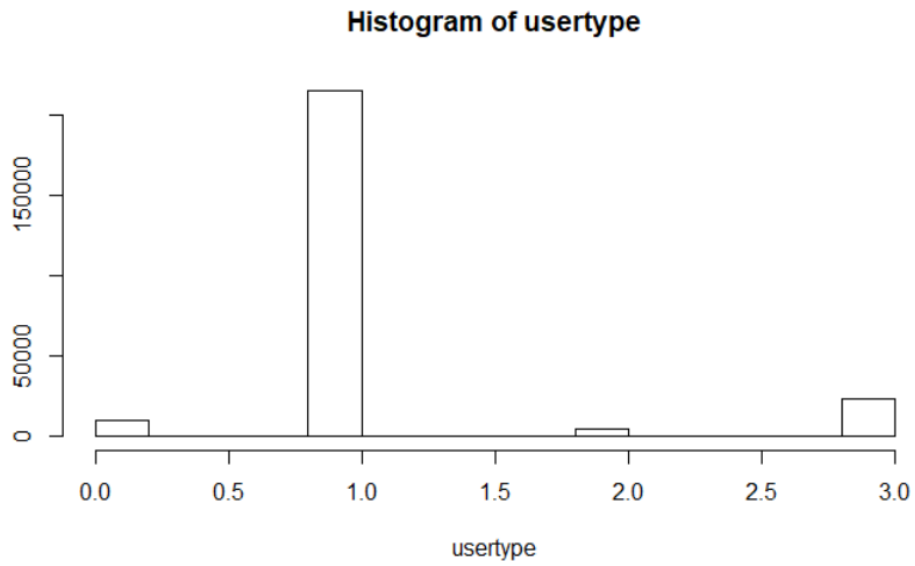


Ilustración 11. Histograma del tipo de usuario de los trayectos (análisis a pequeña escala).

Como se muestra en la Ilustración 11, la mayoría de los usuarios cuentan con un bono anual de bicicletas en BiciMAD (por lo que su 'usertype' es el 1), pero también cabe destacar la presencia, aunque pequeña, de trayectos que han sido realizados por personal de mantenimiento, y que sería conveniente filtrar o considerar de una forma distinta al resto de trayectos antes de integrarlos en el sistema.

Por otro lado, también se analizó los tiempos de los viajes para comprobar cómo de frecuente era que sus valores estuvieran por encima o por debajo de un rango considerado como 'normal' para que la información de un viaje fuera aprovechable y no desvirtuara el análisis y posterior y la creación del modelo.

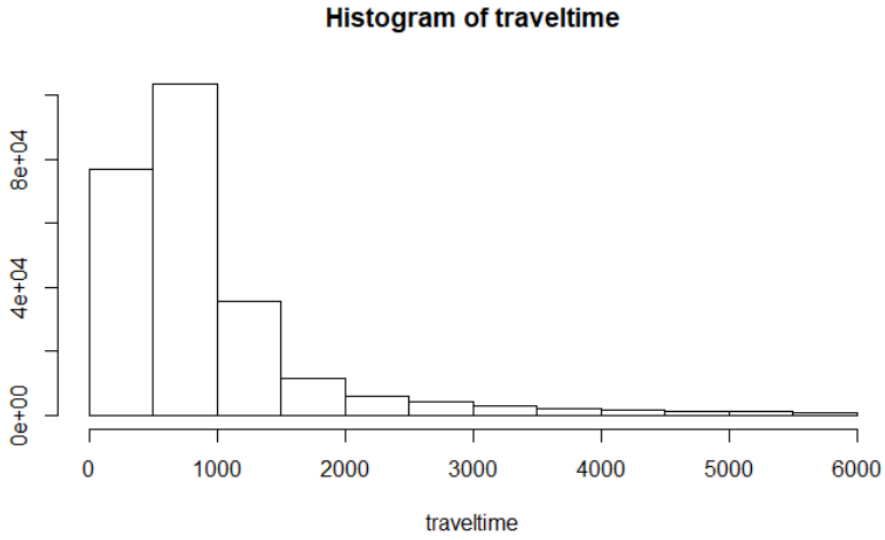


Ilustración 12. Histograma del tiempo de los trayectos (análisis a pequeña escala).

Como se ve en la Ilustración 12, la mayoría de los tiempos están comprendidos en un intervalo normal de tiempos, menor a los 20 minutos de viaje, pero que la cola de la función está muy remarcada y existen valores correspondientes a más de un día. Dado que el objetivo de este proyecto es el análisis de trayectos y la generación de rutas con origen y destino marcados, es necesario filtrar los datos para eliminar aquellos viajes de larga duración.

Como se expresó anteriormente, se puede llevar a cabo un análisis de los tipos de usuarios que utilizan el servicio de bicicletas.

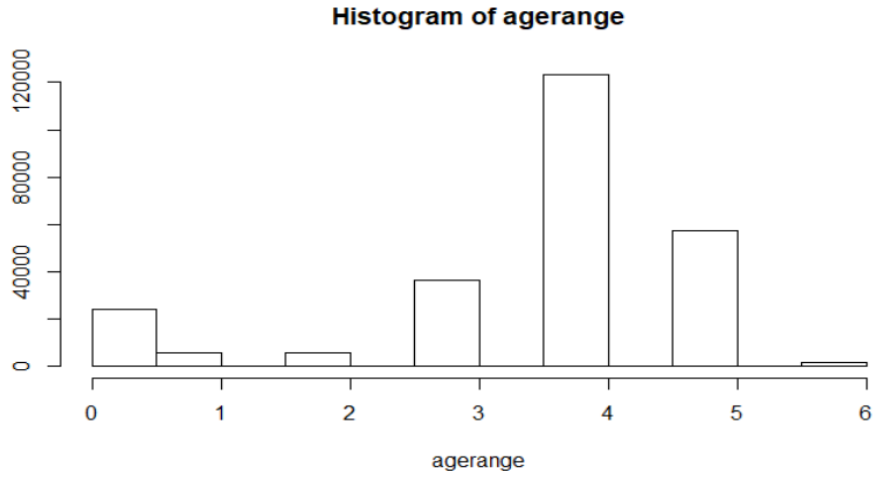


Ilustración 13. Histograma del rango de edad del usuario de los trayectos (análisis a pequeña escala).

Como se muestra en la Ilustración 13, la mayoría de las personas que utilizan el servicio de BiciMAD tienen de 27 a 40 años.

También, considerando información de la hora en la que se ha realizado el viaje, se puede interpretar la causa del uso de la bicicleta, si ha sido como medio de transporte para ir al trabajo/universidad/colegio, por ocio…

También se realizó un análisis del resto de variables que sirvió tanto para conocer qué zonas de Madrid tienen la mayor demanda de bicicletas (conociendo que la estación más utilizada es la situada en la Calle Jesús y María, 36, con la moda de la variable 'unplugstation' o que la mayoría de los usuarios pertenecen a la Zona de Madrid rio/embajadores/mercado s. Miguel, analizando el 'Zipcode'). Por otro lado, se observó que la mayoría de las bicicletas también como destino la Calle Jesús y María, 36, lo que puede significar que se realice una gran cantidad de trayectos que sean para ir y volver al trabajo desde esa estación.

Tras este análisis a pequeña escala de las variables, se trasladó todo este conocimiento adquirido del trabajo con las variables al clúster Big Data de ICAI, con el objetivo de replicar los análisis realizados anteriormente sobre el conjunto total de las variables.

5.1.3 ANÁLISIS A GRAN ESCALA

El análisis a gran escala se realizó con la totalidad de los datos, esto es, 5.333.647 trayectos que se dividen en 3.051.095 viajes en 2017, desde abril hasta diciembre, y 2.282.552 viajes en 2018, desde enero hasta agosto. Estos datos son los disponibles hasta el momento en el que se inició esta parte del proyecto.

El primer objetivo de este análisis a gran escala fue comprobar que las deducciones realizadas sobre la naturaleza de las variables, y las posibles formas de trabajo con éstas descritas anteriormente eran viables tras analizar el total de las variables.

Para esto se repitió la misma estrategia de análisis estadístico realizada sobre las variables, y también se realizaron algunos análisis empezando a cruzar variables para empezar a ver las relaciones que existen entre estas variables.

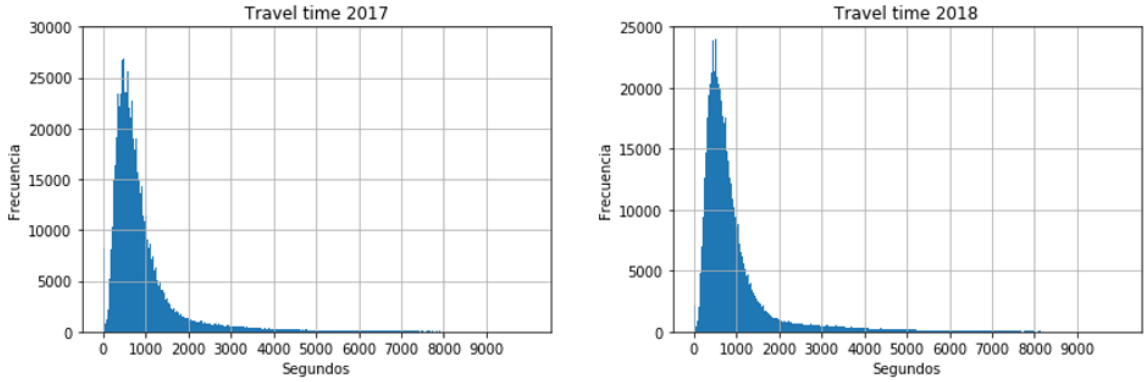


Ilustración 14. Histogramas del tiempo de los trayectos (análisis a gran escala).

Como se afirmó anteriormente, la mayor parte de los trayectos se realizan en un rango inferior a los 2000 segundos y se confirma la existencia de una cola notable en el histograma de los tiempos tras haber añadido todos los datos disponibles en el análisis. También se observa la presencia de tiempos muy pequeños en los trayectos que podría implicar fallos en el sistema de sujeción de las bicicletas o "trayectos" que se generen debido a indecisiones de los usuarios a la hora de elegir la bicicleta. Existen viajes cuyo registro de tiempo es de 0 segundos, mientras que el valor máximo es el de 9.755.123 segundos (7 años y medio). Por

otro lado, hay también datos corruptos, como un registro de tiempo de -3.434 segundos. Independientemente de la naturaleza de estos datos, deben de ser filtrados.

Por otro lado, se identificó que a medida que se analizaban más datos, iban apareciendo una mayor cantidad de errores en éstos. La cantidad de trayectos que tienen variables que han de utilizar el campo por defecto aumenta según aumenta el conjunto analizado.

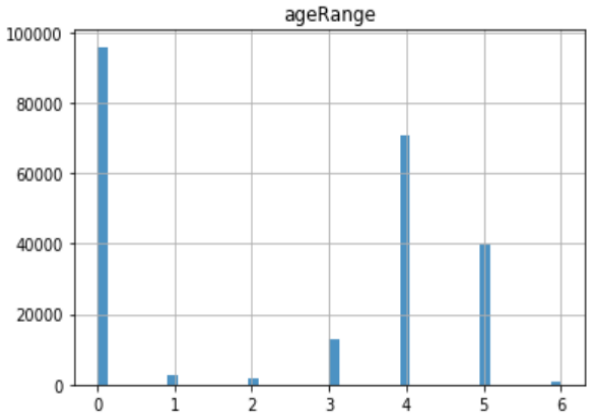


Ilustración 15. Histogramas del rango de edad de los usuarios de los trayectos (análisis a gran escala).

Como se aprecia en la Ilustración 15, hay una gran cantidad de trayectos cuyo campo de rango de edad se ha completado con el valor '0'. Esta situación es un ejemplo de un hecho que acontece en las variables de este tipo; tipos de usuario sin especificar o Zipcodes vacíos son otros casos que se observan al analizar las variables.

Un dato importante para el análisis a gran escala son las estaciones de origen y destino de los trayectos. Dado que la aplicación a desarrollar es el trazado de rutas, conocer la oferta y demanda de bicicletas en las distintas estaciones es importante para poder predecir

el número de bicicletas que habrá en cada estación en el momento en el que se calcule una ruta.

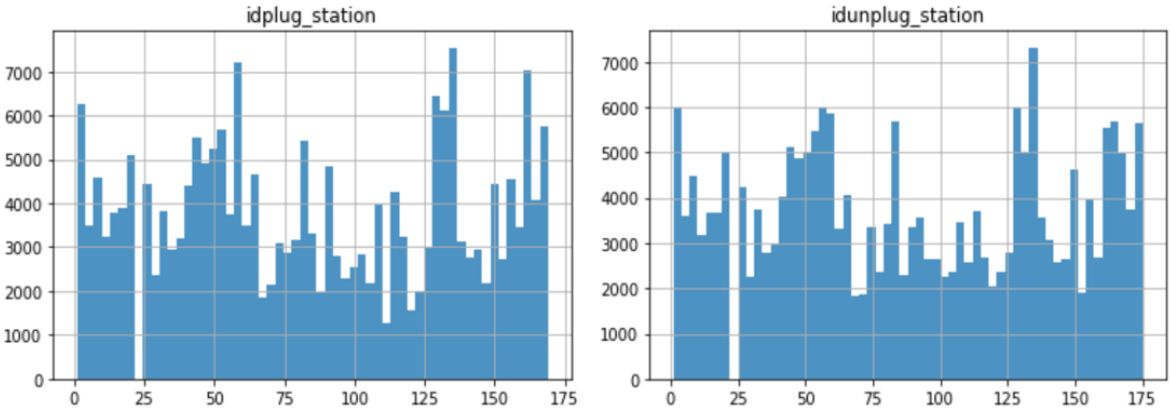


Ilustración 16. Histogramas de la utilización de las estaciones (análisis a gran escala).

En este análisis se aprecia que las estaciones no tienen el mismo uso, además de que hay estaciones que, aunque esté recogida dentro del rango de posibles no se utiliza (ya sea porque ya no existe o porque está deshabilitada), lo que habrá que tener en cuenta a la hora de establecer las rutas.

También en este análisis a gran escala se analizaron variables de forma conjunta, para ver las posibles relaciones entre estas.

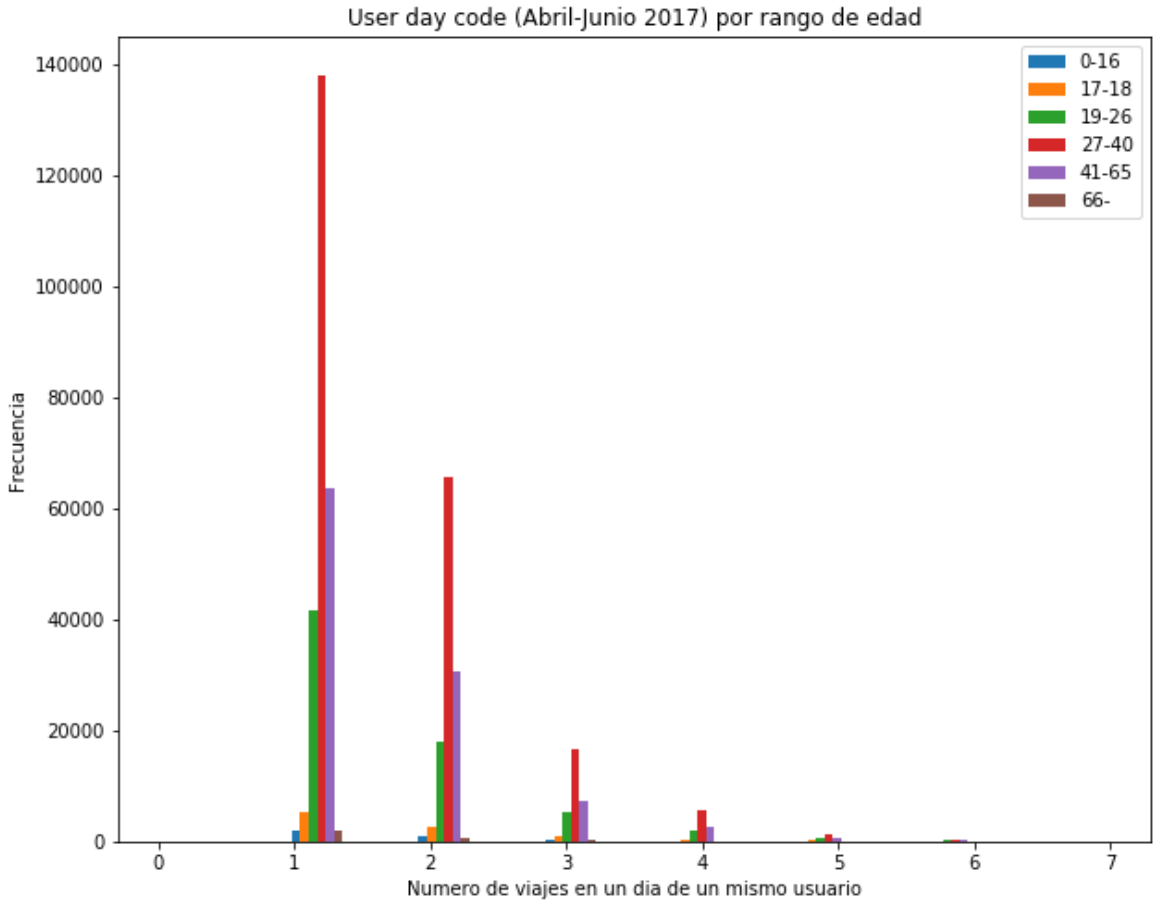


Ilustración 17. Histograma de la cantidad de viajes al día por usuario, dividido según su rango de edad (análisis a gran escala)

Este tipo de análisis conjuntos en variables de este tipo, que segregan a los usuarios por edad y tipo de abono, y también por rango horario, que ayuda al modelo a decidir tramos de rutas según el día y la hora a la que se le solicita un nuevo cálculo de ruta.

Por último, se realizó un análisis de las coordenadas de los trayectos. El objetivo de este análisis no era el de conocer las distintas rutas tomadas ni analizar cada trayecto de forma individual, sino el de conocer el área sobre el que trabaja BiciMAD y poner esta zona de trabajo en relación con los objetivos del proyecto.

Dado que la intención de este proyecto es la de establecer rutas con origen y destino en las estaciones de BiciMAD, hay que considerar un radio de acción a partir del que trabajar, y descartar aquellos trayectos que no cumplan con esta función.

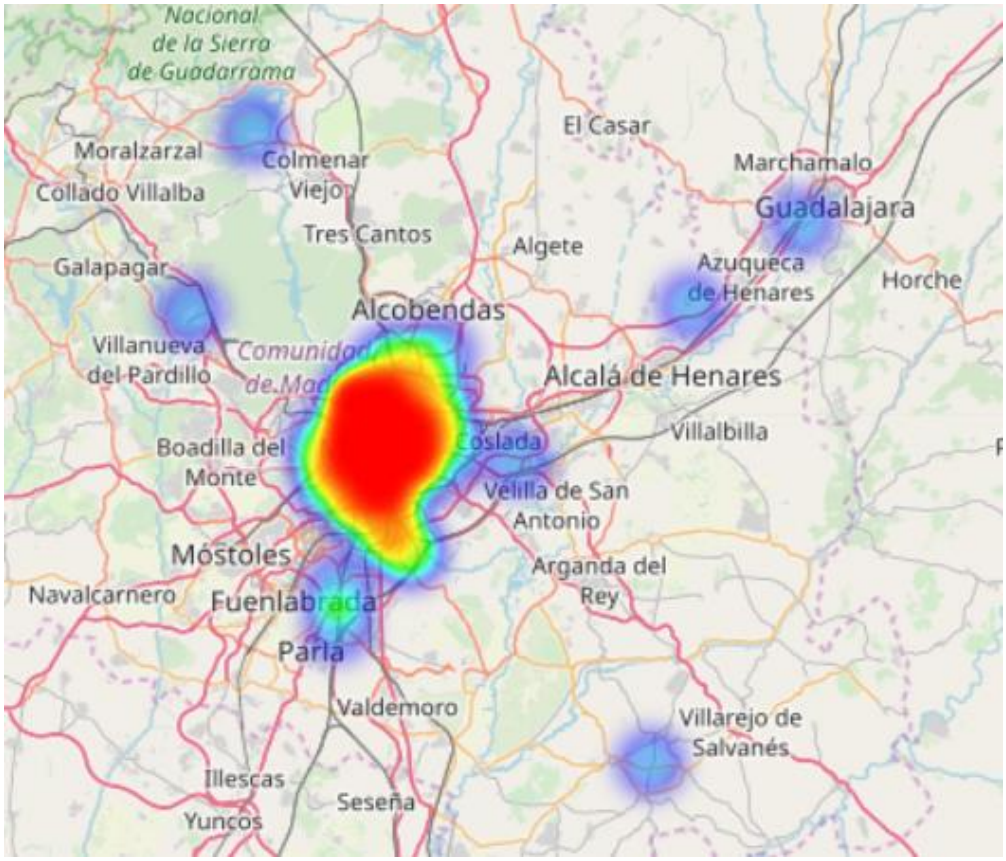


Ilustración 18. Mapa de calor I de trayectos (análisis a gran escala).

Como se comprueba en la Ilustración 18, aunque la gran mayoría de puntos representados se sitúan en el centro de Madrid, existen valores que se alejan del área de trabajo y que, además, son puntos aislados y no una cadena, de forma que no representan una ruta sino fallos a la hora de tomar las medidas.

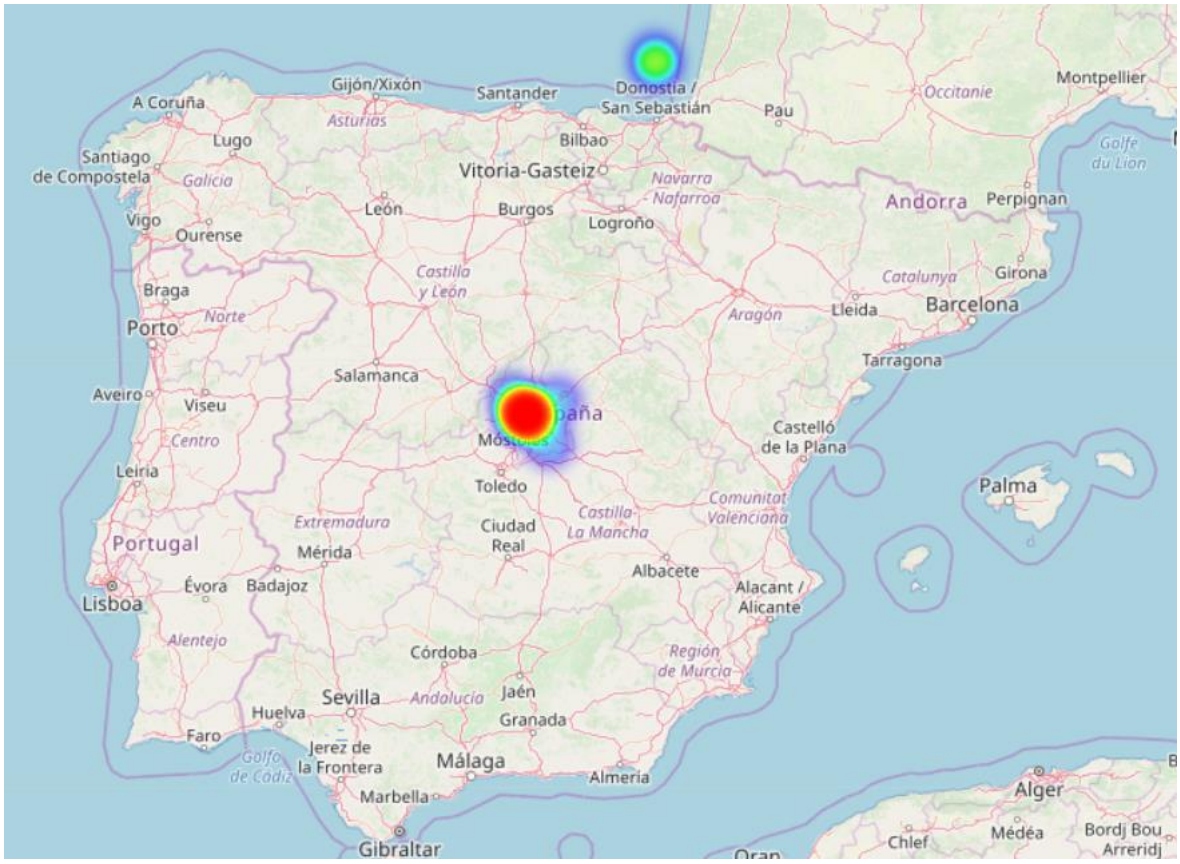


Ilustración 19. Mapa de calor II de trayectos (análisis a gran escala).

En la Ilustración 19 es más notable la existencia de errores. Esto implica la necesidad de que, antes de utilizar estas variables, es necesario filtrarlas y quedarse únicamente con aquellas que representen realmente un trayecto en la ciudad de Madrid.

Las conclusiones a las que nos llevan los análisis de datos explicados previamente son la de una gran necesidad de filtrado de datos para limpiar estos sets de valores que no representan la realidad del proyecto al que nos enfrentamos y otros que claramente son errores del sistema de generación de datos. Además, el conocimiento adquirido con el análisis de estos datos será de vital importancia para todo el trabajo posterior, ya sea con estas mismas variables como con otro tipo de variables cuyo objetivo sea el de complementar este set de datos.

5.2 ACCIDENTES DE BICICLETAS EN MADRID

Como se ha explicado anteriormente, la seguridad es el foco principal en el que se centra este proyecto y que lo diferencia de otras herramientas utilizadas hoy en día. En este proyecto se ha integrado los datos de informes de accidentes de bicicletas ofrecidos en el repositorio del ayuntamiento de Madrid para buscar puntos de peligrosidad en la ciudad de Madrid y, si es conveniente, entrenar al sistema para que trate de evitar esas zonas.

Los datos ofrecidos por el ayuntamiento son .csv en los que se recogen información de los informes que ocurrieron en un determinado año, entre 2011 y 2018. Los datos de mayor importancia de estos archivos son la fecha y hora que recoge el informe del accidente, información de la calle o cruce en el que aconteció, el estado atmosférico o información de las personas involucradas en el accidente (como su sexo o su edad).

Lo más importante que cabe destacar de estos archivos es que no poseen información directa que localicen geográficamente los accidentes en coordenadas, siendo el objetivo principal de esta parte del proyecto inferir las coordenadas a partir de la información de la calle en la que sucedió el accidente.

```
root
 |-- Nombre de la via tratado: string (nullable = true)
 |-- Nombre de la via que cruza: string (nullable = true)
 |-- Longitud en S R  WGS84 (cruce): string (nullable = true)
 |-- Latitud en S R  WGS84 (cruce): string (nullable = true)

root
 |-- Nombre de la v�a: string (nullable = true)
 |-- Literal de numeracion: string (nullable = true)
 |-- Longitud en S R  ETRS89 WGS84: string (nullable = true)
 |-- Latitud en S R  ETRS89 WGS84: string (nullable = true)
```

Ilustración 20. Información de las variables de los accidentes

Para esto se utilizarán dos callejeros de Madrid en los que se recogen información de calles y sus respectivas coordenadas. En uno de ellos habrá una entrada por cada calle con un número de edificio concreto, y en el otro habrá una entrada por cada cruce existente, como se muestra en la Ilustración 20.

Aunque existan dos diccionarios distintos según se necesiten las coordenadas de una calle con su número y un cruce, en los informes de accidentes se identifican tanto un cruce como una calle en el mismo campo, 'LUGAR ACCIDENTE', por lo que habrá que decidir en qué callejero buscar identificando un elemento que sea común a las calles/cruces, pero no al otro.

En la Ilustración 21 se explica el procedimiento a seguir para cada entrada de un accidente:

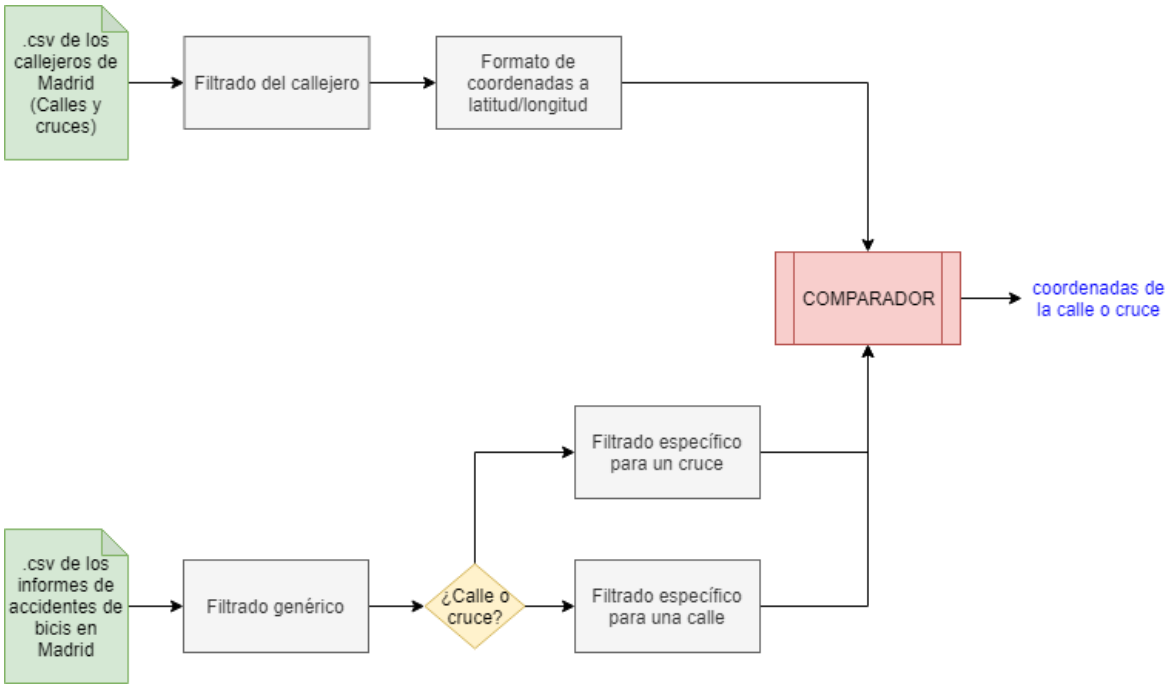


Ilustración 21. Diagrama de la búsqueda de coordenadas para cada Accidente.

El primer paso es integrar en el sistema ambos callejeros para que cada vez que se soliciten las coordenadas de un trayecto se pueda acceder a ellos. Antes de que sean accesibles, como se muestra en el diagrama, su información tiene que ser filtrada y modificada para que concuerden las estructuras tanto de las entradas de los callejeros como las calles recogidas en los .csv de los accidentes.

En el caso del callejero de cruces, aparte de filtrar los nombres también es necesario cruzar las calles para que cada entrada esté repetida, invirtiendo la primera vía del cruce y la

segunda para que luego en la búsqueda del cruce no haya que considerar en qué orden se está introduciendo el cruce ya que, como se explicó anteriormente, el callejero distingue cada vía en un campo distinto, pero el informe del accidente no, y para conservar una estructura común independientemente del accidente es necesario realizar este duplicado.

Tras este filtrado de las calles se procede al cambio del formato de las coordenadas recogidas en los callejeros; mientras que el formato de coordenadas deseado es el de Latitud y Longitud en decimal, estos archivos poseen esta información en grados, minutos y segundos. Tras este cambio en las coordenadas tenemos ambos callejeros listos para trabajar con las calles que se van a introducir.

El primer paso que se lleva a cabo con cada calle que se introduce en el sistema es el de filtrar el nombre para que concuerde lo máximo posible con la estructura utilizada en los callejeros. El principal cambio en el nombre de la calle es el de eliminar términos como ‘Avenida’, ‘Plaza’ o ‘Calle’, que no aparecen en los campos utilizados de los callejeros. Otro cambio también realizado es el de suprimir los artículos y contracciones que aparecen justo antes del nombre de la calle. Este paso justifica el filtrado que hay que realizar en los callejeros, dado que también puede suprimir artículos que sean partes del nombre de la calle.

Por ejemplo, en la calle “CALLE DEL CERRO DE VALDECAHONDE”, interesa suprimir las 2 primeras palabras, que son términos que no aparecen en los callejeros, pero el término “DE” también se filtrará por el sistema dado que es un artículo que, depende de la situación, interesa o no. Aunque en este caso sea parte del nombre de la calle, no importa filtrarlo dado que no posee información útil para asociar el nombre de la calle del accidente con uno del callejero (pero para que este procedimiento funcione correctamente también hay que filtrar el nombre del callejero).

Aparte de esto, en el filtrado también se eliminan calles directamente dado que se puede reconocer simplemente por algún fragmento del texto que la calle no pertenece a la zona de Madrid que interesa (centro de Madrid) sino a los extrarradios, como cuando se informa del kilómetro en el que sucedió el accidente y no el número en la calle.

Tras este primer filtrado genérico toca identificar si el lugar de un accidente está identificado por una calle con número o por un cruce. Tras analizar la estructura de los datos de los accidentes se pudo identificar que el elemento común a todos los cruces, pero que no aparece en ninguna calle con número es un guion (‘-‘) que separa ambas calles del cruce, por lo que este elemento se utilizará tanto de diferenciador para identificar el cruce como de separador para ambas calles del cruce.

El siguiente paso será filtrar por última vez las calles para que tanto los formatos de los callejeros como de las calles sean el mismo. El principal filtrado que se realiza en esta etapa es el del número de la calle, que se hace únicamente sobre las calles porque en los cruces no interesa este campo. En cuanto a los cruces, se separan las calles según el separador determinado anteriormente, se filtra cada una de las calles y se vuelven a unir.

El último paso de este proceso es el de buscar cada calle o cruce en su diccionario correspondiente y sacar el par de coordenadas que identifiquen el lugar del accidente. Este proceso se realiza buscando la calle cuyo nombre sea el que más se acerca al nombre introducido, siendo esta coincidencia mayor del 80% (para evitar que una mínima diferencia, que puede darse debido a fallos humanos, provoque que no se pueda identificar una calle).

El resultado final de este proceso es el mostrado en la Tabla 2:

	FECHA	RANGO HORARIO	CPFA Granizo	CPFA Hielo	CPFA Lluvia	CPFA Niebla	CPFA Seco	CPFA Nieve	TIPO PERSONA	SEXO	Tramo Edad	LATITUD	LONGITUD
0	07/01/2011	DE 15:00 A 15:59	NO	NO	NO	NO	SI	NO	CONDUCTOR	HOMBRE	DE 40 A 44 A�OS	40.44939722222222	-3.602327777777778
1	12/01/2011	DE 18:00 A 18:59	NO	NO	NO	NO	SI	NO	CONDUCTOR	HOMBRE	DE 25 A 29 A�OS	40.45194722222222	-3.683591666666667
2	13/01/2011	DE 9:00 A 9:59	NO	NO	NO	NO	SI	NO	CONDUCTOR	HOMBRE	DE 25 A 29 A�OS	40.44925555555555	-3.717102777777778
3	15/01/2011	DE 10:00 A 10:59	NO	NO	NO	NO	SI	NO	CONDUCTOR	HOMBRE	DE 65 A 69 A�OS	None	None
4	15/01/2011	DE 12:00 A 12:59	NO	NO	NO	NO	SI	NO	CONDUCTOR	HOMBRE	DE 30 A 34 ANOS	None	None

Tabla 2. Accidentes de bicicletas de Madrid con sus coordenadas.

Cada tupla original de los .csv de los accidentes ha sido filtrada para mantener únicamente aquella información de interés para el modelo, cambiando además los campos

que identificaban la calle por su nombre, número y distrito por un set de coordenadas, si el sistema ha sido capaz de sacarlas a través de los diccionarios, o un valor Nulo si no.

5.3 ESTACIONES DE BICIMAD

La necesidad de integrar las estaciones de BiciMAD en el sistema surge por dos motivos. El primero de todos y más obvio, porque en el cálculo de las rutas tanto el origen como el destino ha de ser una de estas estaciones; el segundo es que dentro del set principal de datos de trayectos están marcadas las estaciones tanto de origen como de destino, pero únicamente por su ID, lo que significa que para completar el trayecto hay que cambiar estos identificadores por sus respectivas coordenadas.

El problema surgido con la información de los datos son los reportes de usuarios sobre la dudosa exactitud de los datos recogidos en el Excel oficial ofrecido por el ayuntamiento, lo que supuso la necesidad de una comprobación manual de los datos contrastándolos con una fuente fiable de datos (Google Maps).

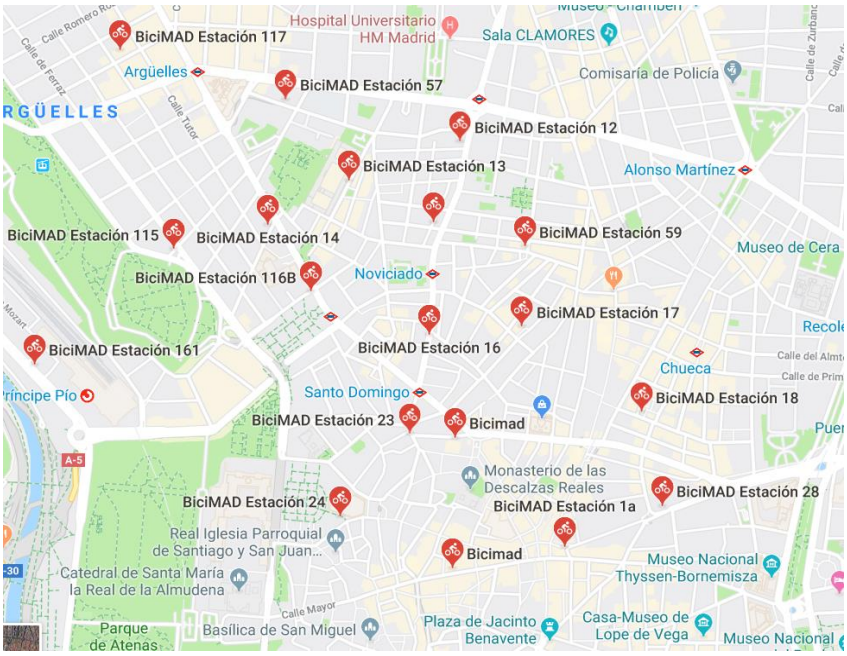


Ilustración 22. Localización de diversas estaciones de BiciMAD en Google Maps.

Tras el la comparación de los datos ofrecidos por el ayuntamiento con los de Google Maps se llegó a la conclusión de que no existía necesidad de crear un set de datos nuevos dado que las únicas diferencias entre la información del ayuntamiento y del buscador web eran de escasos metros, aparte de que algunas estaciones que aparecían duplicadas en el Maps (con sufijos A y B) estaban recogidas como una única estación en la información oficial, por lo que se decidió mantener el set de datos sin alterar su información.

Por último se filtraron los campos del set original para quedarse solamente con aquella información de interés para el modelo posterior, siendo el resultado final el mostrado en la Tabla 3:

	N�mero	Distrito	Barrio	Calle	N° Finca	N°mero de Plazas	LONGITUD	LATITUD
0	001a	01 CENTRO	01-06 SOL	PUERTA DEL SOL, PLAZA, DE LA	1 B	24	-3.702575	40.417016
1	001b	01 CENTRO	01-06 SOL	ALCALA, CALLE, DE	1	24	-3.702462	40.417079
2	2	01 CENTRO	01-05 UNIVERSIDAD	MIGUEL MOYA, CALLE, DE	1	27	-3.705674	40.420580
3	3	07 CHAMBER	07-02 ARAPILES	CONDE DEL VALLE DE SUCHIL, PLAZA, DEL	2	18	-3.707212	40.430322
4	4	01 CENTRO	01-05 UNIVERSIDAD	MANUELA MALASA|A, CALLE, DE	3	27	-3.702674	40.428590
5	5	01 CENTRO	01-04 JUSTICIA	FUENCARRAL, CALLE, DE	106	27	-3.702076	40.428363
6	6	01 CENTRO	01-04 JUSTICIA	HORTALEZA, CALLE, DE	63	24	-3.698409	40.424163
7	7	01 CENTRO	01-04 JUSTICIA	HORTALEZA, CALLE, DE	75	21	-3.697740	40.425216

Tabla 3. Estaciones de BiciMAD tras el filtrado.

5.4 TIEMPO ATMOSFÉRICO

La obtención de información sobre el tiempo atmosférico supone la integración del segundo conjunto de datos relacionados con la seguridad del usuario. La integración de este tipo de información se divide en 2 vías: una mediante una API en tiempo real ofrecida por openweathermap.org, y la segunda es la integración de un histórico de datos mediante WebScrapping obtenido desde la página web de ogimet.com.

5.4.1 API EN TIEMPO REAL

Mediante llamadas a la API de openweathermap, determinando la estación deseada para recibir su información y las unidades de medida (grados Celsius, y velocidad en km/h) recibimos un .json como respuesta a la petición realizada, en la que se proporcionan datos respecto al tiempo atmosférico como las precipitaciones en el momento de realizar la llamada, la velocidad del viento y su dirección, u otros datos respecto a la base que proporciona la información como su ID o su localización. Toda la información recogida en cada llamada es la que se muestra en la Ilustración 23.


```
{'coord': {'lon': -3.7, 'lat': 40.42},
 'weather': [{'id': 501,
   'main': 'Rain',
   'description': 'moderate rain',
   'icon': '10d'}],
 'base': 'stations',
 'main': {'temp': 8.55,
  'pressure': 997,
  'humidity': 100,
  'temp_min': 6.67,
  'temp_max': 10},
 'visibility': 10000,
 'wind': {'speed': 3.6, 'deg': 210},
 'rain': {'1h': 1.02},
 'clouds': {'all': 75},
 'dt': 1556009082,
 'sys': {'type': 1,
  'id': 6399,
  'message': 0.0082,
  'country': 'ES',
  'sunrise': 1555997107,
  'sunset': 1556046053},
 'id': 3117735,
 'name': 'Madrid',
 'cod': 200}
```


Ilustración 23. Llamada a la API de openweathermap.

Por otra parte, se consideró suficiente recibir la información en tiempo real de esta estación (aunque suponga un único punto de medida) dado que su localización es céntrica en la ciudad de Madrid, como se muestra en la Ilustración 24.

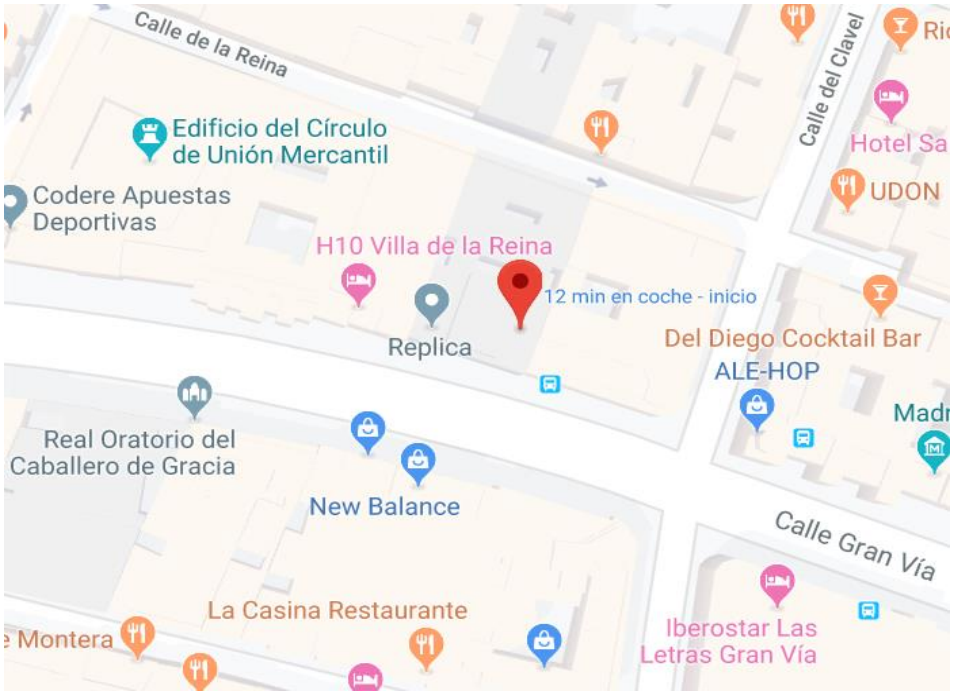


Ilustración 24. Localización geográfica de la estación de medida de openweathermap.

5.4.2 HISTÓRICO DEL TIEMPO ATMOSFÉRICO

El segundo tipo de ingesta de datos sobre el tiempo atmosférico corresponde a la creación de un histórico de datos mediante WebScrapping para relacionar información del tiempo atmosférico con los trayectos y los accidentes de bicicletas. La decisión de realizar la integración de este tipo de datos mediante WebScrapping y no a través de datos ofrecidos por APIs es debido a que no se encontraron sets de datos relacionados con este tipo de información de forma gratuita.

Para extraer la información de la página web se utilizará la librería de Python de BeautifulSoup, mediante la cual se parseará el HTML recibido tras una petición a la página web para poder acceder a distintos elementos indexando según su ID en el HTML, según el tipo de etiqueta, etc.

La estructuración de los datos que se busca adquirir en esta parte de la ingesta es la de los campos de una tabla en la que hay información sobre todas las estaciones disponibles en España.

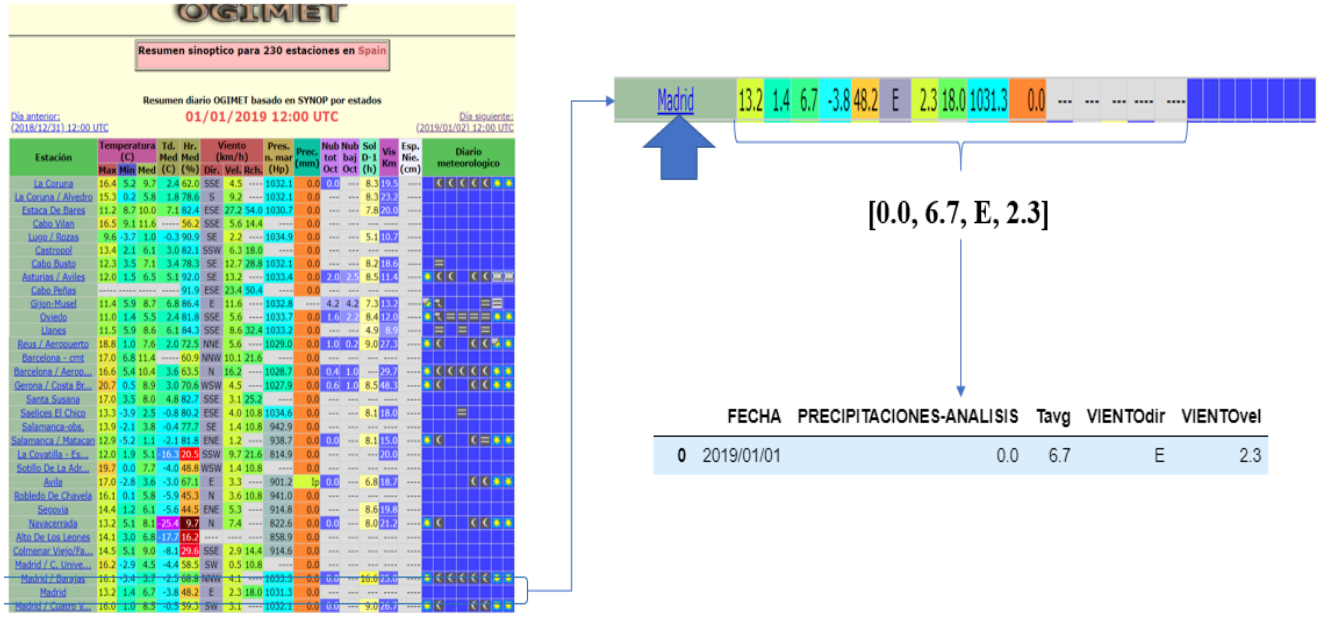


Ilustración 25. Demostración visual de la adquisición de datos por WebScrapping

Como se muestra en la Ilustración 25, la idea principal es acceder indexando por etiquetas a la tupla de la tabla en la que se recoge información de Madrid, para recoger los datos que interesan para etapas posteriores del programa e integrarlos dentro del sistema en una estructura cómoda de usar.

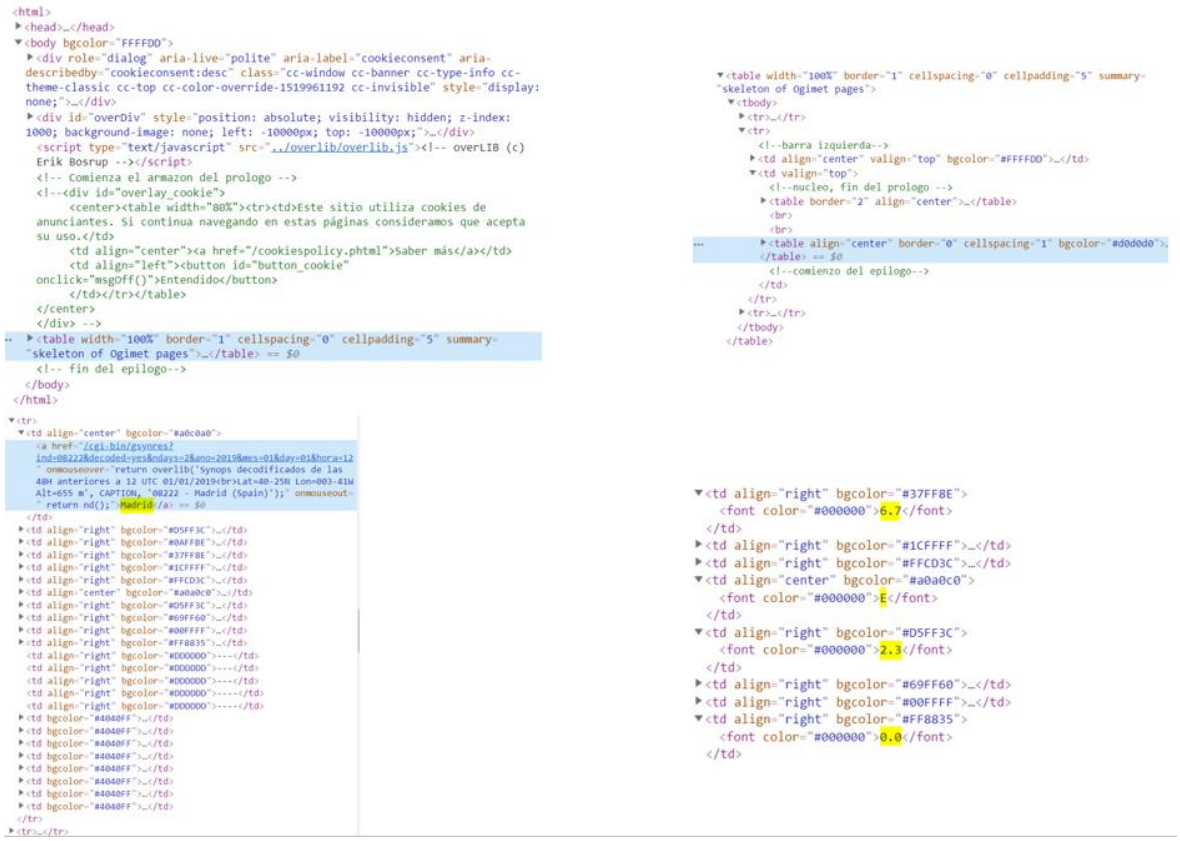


Ilustración 26. HTML de la página web de ogimet.com.

Como podemos observar en la Ilustración 26, el primer paso realizado es acceder al bloque general de la tabla en la que está la información deseada, por lo que mediante el uso de BeautifulSoup filtramos el contenido del HTML para quedarse únicamente con información de las tablas. Tras esto, tenemos que acceder al tercer elemento filtrado, que es la tabla de la que se van a coger los datos de Madrid.

El tercer paso realizado es el de buscar cual de todas las tuplas es la de Madrid, por la que accedemos a los elementos que tengan la etiqueta 'tr' y dentro de este bloque, en el primer elemento con la etiqueta 'td' comprobamos que a tupla es la de Madrid. Por último y tras comprobar que estamos en la zona de la tabla correcta, extraemos todos los sub-bloques de este bloque para quedarnos con los datos que nos interesan mediante la posición que ocupan dentro de la tabla.

El paso final es el de integrar estos datos en el sistema, por lo que se unen todos en una misma variable, añadiendo la información de la fecha de cada dato, y se almacena la variable dentro del clúster.

	FECHA	PRECIPITACIONES-ANALISIS	Tavg	VIENTOdir	VIENTOvel
0	2015/07/02	0.0	29.5	S	14.2
1	2015/07/03	0.0	28.6	SSE	12.7
2	2015/07/04	0.0	30.3	SSE	14.2
3	2015/07/05	0.0	30.5	SSW	8.4
4	2015/07/06	0.0	30.8	WSW	10.1

Tabla 4. Histórico del tiempo atmosférico tras su adquisición.

El resultado final de la creación de este histórico es el mostrado en la Tabla 4, en la que tenemos información de interés para el proyecto dividido por días, desde el 1 de enero de 2011 hasta el 31 de diciembre de 2018.

El principal problema que hay con la adquisición de estos datos mediante la técnica descrita es la necesidad de comprobar en qué tupla de la tabla se está situado antes de acceder a los datos de la fila, debido a que no existe una etiqueta o un identificador único para Madrid, por lo que hay que entrar en el campo en el que indica la estación de recogida de la información. Para solventar este problema se establecieron unos márgenes de búsqueda para buscar en un número reducido de filas en dónde está la información de Madrid.

5.5 TRÁFICO EN TIEMPO REAL

El tráfico en tiempo real es una información de importancia para el proyecto porque tiene relación con los objetivos a la hora de crear una ruta: minimizar el tiempo del trayecto y maximizar la seguridad del usuario.

Los datos en tiempo real del tráfico de Madrid se han integrado en el sistema haciendo uso de la información proporcionada por el repositorio del ayuntamiento en un formato .xml, el cual es accesible mediante una llamada web a un enlace determinado.

La estructura del .xml es el mostrado en la Ilustración 27, en el que se puede apreciar un primer bloque en el que se ofrece información de la hora y fecha en la que se ha actualizado el archivo, y tras esto bloques de información sobre cada uno de los puntos de muestreo del tráfico. Cada muestra se recoge dentro del .xml entre etiquetas del tipo 'pm', por lo que el objetivo será dividir este archivo en estos bloques para proporcionar la información en tiempo real recogida en las coordenadas que se ofrecen como últimos campos de cada bloque (etiquetas 'st_x' y 'st_y').

```
▼<pms>
   <fecha_hora>28/05/2019 15:40:11</fecha_hora>
  ▼<pm>
     <idelem>3409</idelem>
    ▼<descripcion>
       SEPULVEDA Ø118 N-S (CEBREROS-CJAL. FCO. J. JIMENEZ)
     </descripcion>
     <accesoAsociado>240102</accesoAsociado>
     <intensidad>152</intensidad>
     <ocupacion>0</ocupacion>
     <carga>5</carga>
     <nivelServicio>0</nivelServicio>
     <intensidadSat>3000</intensidadSat>
     <error>N</error>
     <subarea>1718</subarea>
     <st_x>436008,175534995</st_x>
     <st_y>4472593,78531503</st_y>
   </pm>
  ▶<pm>...</pm>
  ▶<pm>...</pm>
```

Ilustración 27. Estructura de la información del tráfico en tiempo real.

También podemos observar en la Ilustración 27 que el formato en el que se ofrecen las coordenadas no es el utilizado en este proyecto (latitud/longitud en decimal), por lo que uno de los pasos necesarios a realizar será el cambio de coordenadas. También tenemos un campo de seguridad en el que se informa si los datos son fiables o no. Este campo es el de 'error' y servirá para ignorar ('S') o no ('N') la información del bloque.

Otros datos de interés dentro del bloque son los de 'intensidad', 'ocupación' y 'carga'. El primero de los campos indica el número de vehículos por hora, el segundo de éstos informa del porcentaje de ocupación del punto de control. Por último, la carga es un

parámetro calculado en función de los dos campos anteriormente descritos y de las características de la vía en la que se tome la muestra.

Teniendo en cuenta todos los campos y la diversa información recogida en cada bloque sobre el tráfico, el trabajo a realizar con este archivo web es el mostrado en la Ilustración 28:

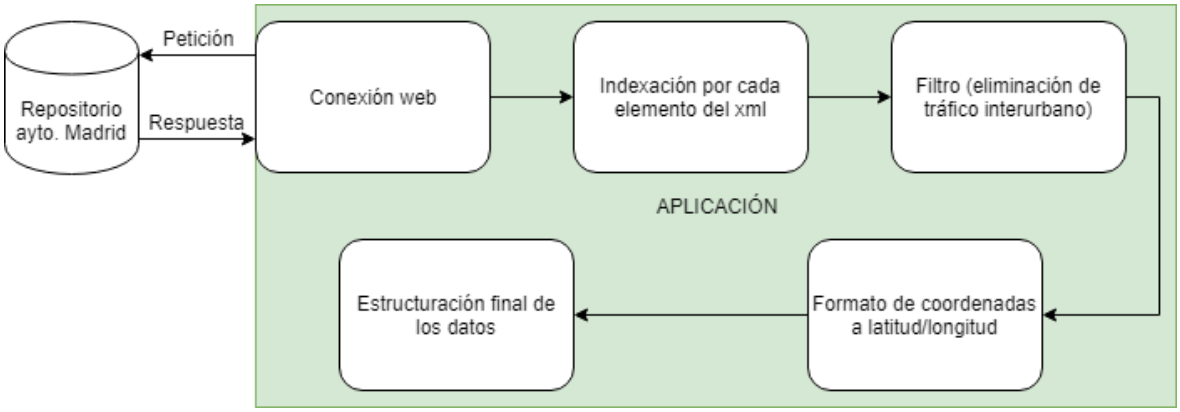


Ilustración 28. Diagrama del tratamiento del .xml con información del tráfico en tiempo real.

Primero de todo se realizará una petición web a la dirección del ayuntamiento que nos devolverá el archivo XML, se abrirá y se parseará para poder acceder a sus campos.

Tras esto se accederá a cada bloque recogido del xml para acceder a sus datos. La primera etiqueta que se buscará será la de ‘descripción’, puesto que si un bloque no posee esta etiqueta significa que lo podemos descartar, dado que el bloque tiene información de tráfico interurbano, y como ya se ha explicado anteriormente, no interesa. Tras haber validado el bloque, se cogerá toda la información para almacenarla en el clúster.

Antes de guardar la información, y como se ha mencionado anteriormente en este capítulo, hay que cambiar el formato de las coordenadas para que concuerden con el formato utilizado en el resto de las aplicaciones del proyecto, se cambiaran del formato UTM, con posición localizada en ’30-T’ para Madrid, a latitud y longitud en decimal mediante el uso de la librería ‘utm’ de Python.

Por último, se cambiará el formato de las variables para que concuerden con la información que representan (por ejemplo, la información de intensidad y carga a Float) para que estén disponibles para ser utilizados por el resto de integrantes del proyecto. La estructura final de estos datos se muestra en la Tabla 5.

	IDELEM	DESCRIPCION	ACCESOASOCIADO	INTENSIDAD	OCUPACION	CARGA	NIVELSERVICIO	INTENSIDADSAT	ERROR	SUBAREA	COORDX	COORDY
0	3409	SEPULVEDA Ø118 N-S (CEBREROS-CJAL. FCO. J. JIM...	240102	189	2	8	0	3000	N	1718	40.4015	-3.7541
1	4739	CJAL. FCO. J. JIMENEZ Ø126 E-O (BERLANAS-SEPUL...	240101	379	3	15	0	3000	N	1718	40.3997	-3.75372
2	4740	CJAL. FCO. J. JIMENEZ Ø86 O-E (F. CALVO-ALHAMBRA)	240604	254	2	10	0	3000	N	1712	40.4005	-3.74572
3	4741	CJAL. FCO. J. JIMENEZ Ø76 E-O (HURTUMPASCUAL-A...	240602	212	1	8	0	3000	N	1712	40.4009	-3.74357

Tabla 5. Estructuración final de los datos de tráfico en tiempo real.

5.6 POLUCIÓN EN LA COMUNIDAD DE MADRID

Otra de las áreas que se desea cubrir en este proyecto es la de la relación entre la polución y el uso de medios de transporte no contaminantes. Con las medidas anticontaminación llevadas a cabo por el gobierno de Madrid durante el mandato que terminó en 2019, como por ejemplo 'Madrid Central', la contaminación se puede haber distribuido de una forma no uniforme a lo largo de la ciudad de Madrid.

Uno de los objetivos que se busca conseguir con la integración de estos datos es el ver una posible relación entre las rutas tomadas por los usuarios de BiciMAD y la distribución de esta contaminación, mediante el uso de históricos de polución. El otro objetivo es el de ofrecer al usuario una ruta que pase por zonas menos contaminadas de Madrid; que sea más sana para el ciclista.

Para esta parte del proyecto se contará con la información ofrecida en el repositorio del ayuntamiento captada a través de estaciones de contaminación distribuidas a lo largo de Madrid.

5.6.1 ESTACIONES DE POLUCIÓN

El primer paso de todos es la integración de la información relacionada con las estaciones de polución, en la que se proporcionan datos sobre la localización de las estaciones, su identificador (utilizado más adelante tanto en los históricos como en la información en tiempo real), o el tipo de información sobre polución recogida, como se muestra en la Tabla 6.

N◆MERO	ESTACI◆N	DIRECCI◆N	LONGITUD	LATITUD	ALTITUD	TIPO ESTACION	NO2	SO2	CO	...	BTX	HC	UV	VV	DV	TMP	HR
4	Pza. de Espa◆a	Plaza de Espa◆a	-3.7122472222222225	40.423852777777775	635	UT	X	X	X	...	None	None	None	X	X	X	X
8	Escuelas Aguirre	Entre C/ Alcal◆ y C/ O◆ Donell	-3.6823194444444445	40.42156388888888	670	UT	X	X	X	...	X	X	None	None	None	None	None
11	Avda. Ram◆n y Cajal	Avda. Ram◆n y Cajal esq. C/ Pr◆ncipe de Vergara	-3.6773555555555553	40.451475	708	UT	X	None	None	...	X	None	None	None	None	None	None
16	Arturo Soria	C/ Arturo Soria esq. C/ Vizconde de los Asilos	-3.6392333333333333	40.44004722222222	693	UF	X	None	X	...	None	None	None	None	None	None	None

Tabla 6. Estaciones de polución del ayuntamiento de Madrid.

Ha sido necesario alterar los datos de las coordenadas ofrecidas en el archivo original de las estaciones debido a que su formato era el de grados, minutos y segundos, mientras que el utilizado en este proyecto es el decimal.

5.6.2 INFORMACIÓN DE POLUCIÓN

El siguiente paso en esta parte del proyecto es la adquisición de la información disponible sobre polución. Para esto se cuentan con datos tanto en tiempo real, disponibles de varias fuentes según las necesidades en el momento concreto en el que se solicitan los datos, como un histórico de polución en Madrid ofrecido por el ayuntamiento.

Tal como se muestra en la Ilustración 29, dependiendo de la procedencia de los datos y del tipo de los mismos, se realizarán procesos distintos con estos. También es importante apreciar que, dado que el formato en el que se ofrecen todos los datos desde el ayuntamiento, tanto históricos como datos en tiempo real, son .txt, por lo que se utilizarán las mismas funciones para extraer la información relevante e integrarla en el sistema.

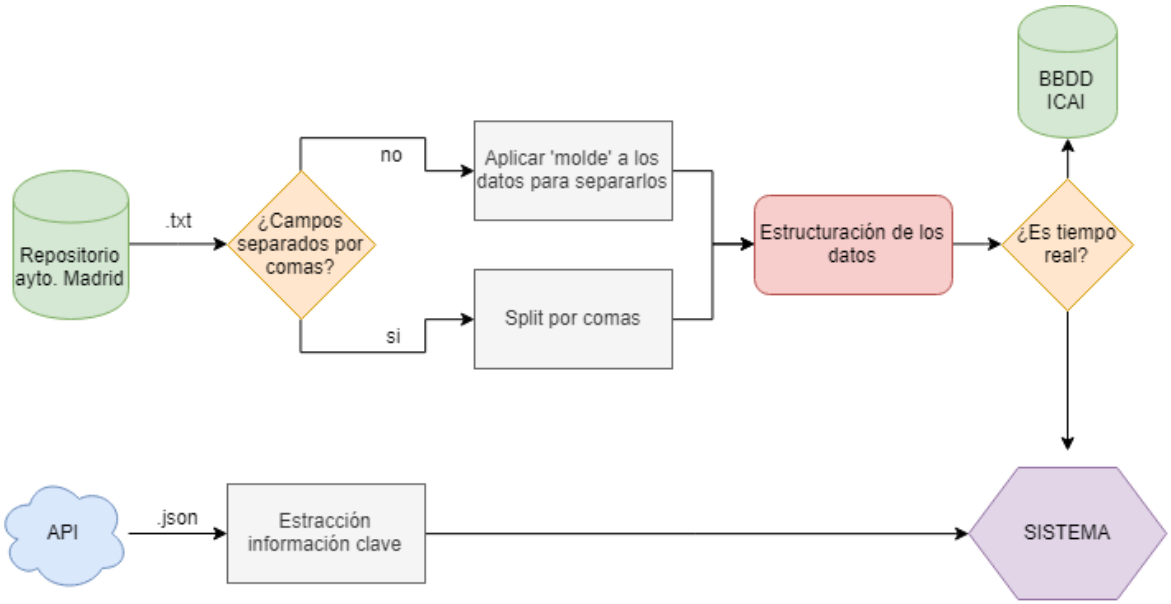


Ilustración 29. Diagrama de la integración de datos de polución.

Tal como se muestra en la Ilustración 29, los datos del repositorio del ayuntamiento se reciben en formato .txt, por lo que el primer paso será extraer las línea en un formato de String para separar cada uno de los datos y asignarlos a su campo correspondiente (la información sobre los campos de los datos viene dada en un archivo aparte por el ayuntamiento).

El principal problema que tiene este set de datos es que los archivos de los históricos tienen 2 formatos distintos: Desde enero de 2011 hasta octubre de 2017 los datos están almacenados sin utilizar ningún tipo de separador, por lo que fue necesario parsear las líneas de los .txt conociendo exactamente la longitud que tenía cada uno de los datos, y dividiendo el String con este "molde". A partir de noviembre de 2017 (incluyendo los datos en tiempo real) los datos vienen divididos por comas (','), por lo que con un simple split del String por este carácter se pueden dividir los datos de cada línea.

El siguiente paso realizado con este tipo de datos es el de unir toda la información de las distintas líneas que se han parseado, pero en este caso no es necesario añadir una marca de la fecha de cada dato debido a que esta información ya está disponible a través de tres campos: los de año, mes y día.

Tras esto, si los datos parseados son parte del histórico, se almacenarán en el clúster, y si no podrán integrarse para utilizarlos en otras partes del proyecto.

En la Tabla 7 podemos ver cuál es la disposición final de los datos de polución del ayuntamiento, independientemente de si son históricos o datos en tiempo real.

MUNICIPIO	ESTACION	MAGNITUD	TECNICA	PERIODO-ANALISIS	ANHO	MES	DIA	00	...	14	15	16	17	18	19	20	21	22	23
079	004	01	38	02	2019	04	10	00003	...	00004	00003	00000	00000	00000	00000	00000	00000	00000	00000
079	004	06	48	02	2019	04	10	000.3	...	000.4	000.4	00000	00000	00000	00000	00000	00000	00000	00000
079	004	07	08	02	2019	04	10	00002	...	00012	00012	00000	00000	00000	00000	00000	00000	00000	00000
079	004	08	08	02	2019	04	10	00015	...	00032	00034	00000	00000	00000	00000	00000	00000	00000	00000
079	004	12	08	02	2019	04	10	00017	...	00050	00053	00000	00000	00000	00000	00000	00000	00000	00000
079	008	01	38	02	2019	04	10	00008	...	00009	00010	00000	00000	00000	00000	00000	00000	00000	00000
079	008	06	48	02	2019	04	10	000.1	...	000.2	000.1	00000	00000	00000	00000	00000	00000	00000	00000
079	008	07	08	02	2019	04	10	00004	...	00011	00008	00000	00000	00000	00000	00000	00000	00000	00000
079	008	08	08	02	2019	04	10	00039	...	00047	00035	00000	00000	00000	00000	00000	00000	00000	00000

Tabla 7. Datos de polución del ayuntamiento de Madrid.

La primera información mostrada en la tabla es la relacionada con la estación en la que se han adquirido los datos y la fecha en la que se han captado. Tras eso, se muestran 24 campos (de '00' a '23', una muestra cada hora del día) referentes al nivel de polución.

Por otra parte, como también se muestra en el diagrama de la Ilustración 29, se ha utilizado la API en tiempo real de aqicn.org, que nos proporciona información de polución en una única estación en Madrid. La API ofrece información sobre los niveles de contaminación de distintos gases, y la interpretación de estos niveles es tal como se muestra en la Tabla 8:

AQI	Air Pollution Level	Health Implications	Cautionary Statement (for PM2.5)
0 - 50	Good	Air quality is considered satisfactory, and air pollution poses little or no risk	None
51 -100	Moderate	Air quality is acceptable; however, for some pollutants there may be a moderate health concern for a very small number of people who are unusually sensitive to air pollution.	Active children and adults, and people with respiratory disease, such as asthma, should limit prolonged outdoor exertion.
101-150	Unhealthy for Sensitive Groups	Members of sensitive groups may experience health effects. The general public is not likely to be affected.	Active children and adults, and people with respiratory disease, such as asthma, should limit prolonged outdoor exertion.
151-200	Unhealthy	Everyone may begin to experience health effects; members of sensitive groups may experience more serious health effects	Active children and adults, and people with respiratory disease, such as asthma, should avoid prolonged outdoor exertion; everyone else, especially children, should limit prolonged outdoor exertion
201-300	Very Unhealthy	Health warnings of emergency conditions. The entire population is more likely to be affected.	Active children and adults, and people with respiratory disease, such as asthma, should avoid all outdoor exertion; everyone else, especially children, should limit outdoor exertion.
300+	Hazardous	Health alert: everyone may experience more serious health effects	Everyone should avoid all outdoor exertion

Tabla 8. Niveles de contaminación en la API en tiempo real de aqicn.org

5.7 ADAPTACIÓN FINAL DE LOS DATOS

El último trabajo llevado a cabo en este proyecto ha sido la adaptación final de los datos que se han ido tratando a lo largo de este capítulo. Las adaptaciones realizadas en esta parte del proyecto son de distinto carácter:

Primero de todo se adaptaron las marcas de fechas de los datos para que todos los archivos tuvieran una misma estructura (AAAA/MM/DD) y se pudieran unir datos directamente.

Otro dato alterado fueron las marcas de edad presentes tanto en los datos de trayectos como en los de accidentes. Se utilizará el formato definido en el capítulo 5.1, concretamente en el subapartado 'AGERANGE'.

Por último, se alteraron los valores de las variables de los distintos sets de datos para utilizar los valores que más faciliten el procesamiento posterior de los datos. De este modo, en campos como las fechas se utilizan Strings, edades o ID numéricos Integers, y en datos de medidas tomadas y coordenadas se formatean como Floats.

Capítulo 6. ANÁLISIS DE RESULTADOS

Tras la realización de este proyecto de ingesta de datos para un sistema Big Data es necesario el análisis de todos los resultados y conclusiones que se han obtenido.

Primero de todo, los sets de datos proporcionados por el ayuntamiento de Madrid en su repositorio abierto dificultan la creación de un sistema sólido de captación de datos debido a que estos archivos carecen de estructuras cómodas para el trabajo realizado. Póngase como ejemplo de esto el trabajo realizado en la integración de los datos de Polución, teniendo que tomar la decisión de forma dinámica de dividir manualmente los datos o utilizar los separadores pertinentes debido a que, dependiendo del archivo, tienen una estructura u otra.

Por otro lado, el enfoque del proyecto a la seguridad supuso la integración de datos propios de accidentes, y estos datos no están preparados para utilizarse en el entorno en el que se están integrando, lo que supuso la programación de un sistema de obtención de coordenadas a partir de nombres de calles que tenía que extraer los datos del entorno hdfs para trabajar con ellos, y que supuso un coste computacional más alto que el normal en un contexto Big Data.

En cuanto a la utilización de WebScrapping en el proyecto, supone una herramienta de gran utilidad, puesto que proporciona la posibilidad de adquirir datos sin que esto sea directamente a través de una API o de archivos. Los resultados de la utilización de esta herramienta son la creación de un histórico que proporciona información importante en el proyecto que ha sido obtenida utilizando este método debido a la imposibilidad de encontrar información de históricos de tiempo proveniente de otros tipos de fuentes. El problema de utilizar esta técnica de obtención de datos es que los programas creados para una página web no pueden ser utilizados para un análisis de otra web, aunque la estructura de ambas webs sea similar, debido a la necesidad de indexar por etiquetas y por elementos propios del HTML o XML como los IDs o las clases.

Otro punto que considerar es la en la integración de datos en tiempo real se llegan a dos resultados claramente diferenciados:

1. Por un lado, tenemos los datos en tiempo real claramente definidos, que son los provenientes de APIs creadas con esta intención como lo son las de openweathermap para el tiempo, o aqicn para la polución, que suponen una fuente de datos rápida (pues este es el objetivo de utilizar este tipo de APIs) pero cuya información es muy escasa, pues en ambos casos se dan medidas tomadas en un único punto de referencia.
2. Por el otro lado, tenemos los datos en 'pseudo tiempo real' provenientes del ayuntamiento. Es necesario utilizar un nombre diferenciador del otro set de datos dado que el tratamiento de estas fuentes de datos es más parecido al utilizado en la creación de los históricos que en una API en tiempo real. La utilización de este tipo de datos supone una mayor carga computacional que ha de llevarse a cabo cada vez que se solicite una nueva ruta, lo que al final supone proveer más información a costa de la velocidad.

Para resumir este punto y como se ha ido explicando en el Capítulo 5 relativo a los datos en tiempo real, la utilización de un tipo de fuente sobre otra vendrá de la preferencia de una mayor cantidad de información o de la necesidad de más velocidad.

Por último, mostrar información de correlaciones de algunos datos adquiridos en este proyecto. Este análisis final se ha realizado con datos correspondientes a los meses en los que hay información de los trayectos de bicicletas. Aunque hay variables que se puede presuponer que tienen algún tipo de relación, como podría ser la cantidad de trayectos y la temperatura o las precipitaciones, puede surgir que tras un análisis posterior estas presunciones no son ciertas. Por otro lado, también parece normal suponer que la relación entre trayectos de bicicletas al día y el número de accidentes de bicis al día es alta, pero realizar este análisis con pocos datos (de accidentes en este caso) puede llevar a que la relación obtenida se aleje de lo que se supone como normal.

En la Tabla 9 se muestran correlaciones obtenidas de los datos, en los que se puede ver que, aparte de relaciones directamente proporcionales que son obvias (como temperaturas) el resto de las variables no tienen relaciones tan claras.

	ViajesDia	PRECIPITACIONES-ANALISIS	Tavg	Tmax	Tmin	VIENTOvel	AccidentesBiciDia
ViajesDia	1.000000	-0.315534	0.159615	0.197539	0.095642	-0.275115	0.272331
PRECIPITACIONES-ANALISIS	-0.315534	1.000000	-0.164087	-0.190513	-0.054793	0.244156	-0.096736
Tavg	0.159615	-0.164087	1.000000	0.977044	0.964351	-0.132932	0.215969
Tmax	0.197539	-0.190513	0.977044	1.000000	0.905003	-0.223691	0.232747
Tmin	0.095642	-0.054793	0.964351	0.905003	1.000000	-0.015785	0.183124
VIENTOvel	-0.275115	0.244156	-0.132932	-0.223691	-0.015785	1.000000	-0.131351
AccidentesBiciDia	0.272331	-0.096736	0.215969	0.232747	0.183124	-0.131351	1.000000

Tabla 9. Correlaciones de accidentes, trayectos y datos atmosféricos.

Capítulo 7. CONCLUSIONES Y TRABAJOS FUTUROS

7.1 CONCLUSIONES

Como conclusión, la ingesta de datos en Big Data es un proceso necesario en todo sistema, pero que a su vez está poco evolucionado y necesita de un amplio conocimiento de cada fuente que se quiere integrar. Esto lleva a que los sistemas de ingesta de datos tengan que crearse específicamente para cada tipo de fuente que se necesite, y que la posibilidad de reutilizar una herramienta sin tener que alterar su uso para una fuente de datos concreta es muy escasa. Esto también lleva a la dificultad de escalar un sistema de ingesta, debido a que están creados de forma concreta para un uso, y la creación de elementos para ingesta no puede ser independiente a la fuente. Esto lleva a que la elaboración de herramientas de este ámbito han de estar enfocados a elementos concretos, y no para la utilización en un sector o en una estructura genérica.

En este proyecto de ingesta de datos se ha llevado a cabo una adquisición manual de los archivos necesarios para crear la base necesaria sobre la que se apoyan el resto de las partes del proyecto. También se ha necesitado realizar un procesamiento de un carácter artesanal debido a las causas explicadas en el análisis de resultados.

La conclusión de este proyecto es que se ha conseguido crear archivos de datos con información diversa para nutrir el resto del proyecto del soporte necesario para que realicen su función.

7.2 TRABAJOS FUTUROS

Existe la posibilidad de llevar a cabo futuros desarrollos relacionados con este proyecto, que se podrían dividir en dos ramas fundamentales:

7.2.1 AGREGACIÓN DE FUENTES DE DATOS

La primera de ellas sería la agregación de otros tipos de información para proporcionar un mayor abanico de posibilidades de elección al modelo posterior, así como la posibilidad de visualizar información de interés para los ciclistas en la ciudad de Madrid

Una posibilidad podría ser explotar la herramienta del WebScrapping para conseguir información utilizando páginas web de referencia de noticias, consiguiendo de este modo datos sobre posibles accidentes, cortes de carreteras o eventos próximos.

También, aprovechando el gran uso del repositorio de datos del ayuntamiento en este proyecto, es posible la integración de otros datos disponibles, como información en tiempo real del estado de las estaciones o estudiar la posibilidad de añadir el histórico de accidentes de todo tipo de vehículos.

Por último, una de las intenciones de este proyecto era la integración de la API de Twitter para sacar noticias en tiempo real dadas por los propios usuarios de la red social, por lo que se podría desarrollar en el futuro un análisis de redes sociales para conseguir información en tiempo real de eventos que puedan alterar el cálculo de una ruta óptima.

7.2.2 DEPURACIÓN DEL CÓDIGO ORIGINAL

Por otro lado, existen partes del código original que podrían ser alterados para que, aunque el resultado original sea semejante al obtenido tras la realización del proyecto actual, se utilice la infraestructura Big Data con la que cuenta ICAI de un modo eficiente.

Ejemplo claro de esta situación es el modo en el que se sacan las coordenadas a partir de los accidentes, en el que se podría separar en dos partes claramente diferenciadas el filtrado de los campos con la búsqueda de coordenadas mediante el uso de diccionarios.

Capítulo 8. REFERENCIAS

[1] «IBM Big Data Hub,» [En línea]. Available: https://www.ibmbigdatahub.com/infographic/where-does-big-data-come.

[2] S. Sinha, «Edureka,» [En línea]. Available: https://www.edureka.co/blog/hadoop-tutorial/.

[3] M. A. e. al, «A Data-Driven Knowledge Acquisition System: An End-to-End Knowledge Engineering Process for Generating Production Rules,» de IEEE Access.

[4] P. T. M. B. A. Devlin, «An architecture for a business and information system,» de IBM Systems Journal.

[5] H. G. a. S.-T. L. Sanjay Ghemawat, «Google,» 2003. [En línea]. Available: https://static.googleusercontent.com/media/research.google.com/es//archive/gfs-sosp2003.pdf.

[6] [En línea]. Available: https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html.

[7] M. W. A. B. M. W. Zoiner Tejada, «Microsoft Azure,» [En línea]. Available: https://docs.microsoft.com/en-us/azure/architecture/data-guide/big-data/.

[8] E. Wirth, «MathSection,» 12 Diciembre 2018. [En línea]. Available: https://mathsection.com/how-google-maps-calculates-the-shortest-route/.

[9] R. Gutman, «Quora,» 28 Octubre 2017. [En línea]. Available: https://www.quora.com/How-does-the-algorithm-of-Google-Maps-work.

[10] «Ciclosfera,» 14 Noviembre 2016. [En línea]. Available: https://www.ciclosfera.com/madbike-app-bicimad/.

[11] D. Hassan, «A Text Mining Approach for Evaluating Event Credibility on Twitter,» de 27th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), Paris, 2018.

[12] R. A. C. N. R. A. A. Kadadi, «Challenges of data integration and interoperability in big data,» de IEEE International Conference on Big Data (Big Data), Washington, DC, 2014.

[13] L. P. T. C. X. W. S. Liu, «Research on multi-source heterogeneous data collection for the Smart City public information platform,» de IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, 2016 .

[14] M. M. F. Cordeiro, «Mining the Twitter Stream: Unravel Events, Interactions, and Communities in Real-Time,» de 17th IEEE International Conference on Mobile Data Management (MDM), Porto , 2016.

[15] S. Ahuja, «Discovering significant news sources in Twitter,» de IEEE International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), Kolkata, 2015.

[16] A. M. B. N. Dhavase, «Location identification for crime & disaster events by geoparsing Twitter,» de International Conference for Convergence for Technology, Pune, 2014.

[17] S. Kempe, «Dataversity,» 23 August 2012. [En línea]. Available: https://www.dataversity.net/a-short-history-of-data-warehousing/#.

[18] SAS, «SAS,» [En línea]. Available: https://www.sas.com/es_ar/insights/data-management/what-is-etl.html.

[19] J. Baker, «Medium,» 1 October 2017. [En línea]. Available: https://becominghuman.ai/getting-started-with-building-realtime-api-infrastructure-a19601fc794e.

[20] J. Williams, «Prompt Cloud,» 24 September 2018. [En línea]. Available: https://www.promptcloud.com/blog/scrape-twitter-data-using-python-r/.

[21] M. Ebrahim, «Like Geeks,» 5 December 2017. [En línea]. Available: https://likegeeks.com/es/web-scraping-beautiful-soup-y-selenium/#Que-es-Web-Scraping-con-Python.

[22] V. Paruchuri, «DataQuest,» 17 November 2016. [En línea]. Available: https://www.dataquest.io/blog/web-scraping-tutorial-python/.

[23] L. Richardson, «Beautiful Soup,» [En línea]. Available: https://beautiful-soup-4.readthedocs.io/en/latest/.