An ETL-like platform for the processing of mobility data

Maxime Masson, Cécile Cayèré, Marie-Noëlle Bessagnet, Christian Sallaberry, Philippe Roose, Cyril Faucher

An ETL-like platform for the processing of mobility data

Maxime Masson
maxime.masson@univ-pau.fr
LIUPPA, E2S, University of Pau and Pays Adour (UPPA)
France

Cécile Cayèré
cecile.cayere1@univ-lr.fr
L3i, La Rochelle University
France

Marie-Noëlle Bessagnet
marie-noelle.bessagnet@univ-pau.fr
LIUPPA, E2S, University of Pau and Pays Adour (UPPA)
France

Christian Sallaberry
christian.sallaberry@univ-pau.fr
LIUPPA, E2S, University of Pau and Pays Adour (UPPA)
France

Philippe Roose
philippe.roose@univ-pau.fr
LIUPPA, E2S, University of Pau and Pays Adour (UPPA)
France

Cyril Faucher
cyril.faucher@univ-lr.fr
L3i, La Rochelle University
France

ABSTRACT

In this article, we introduce a novel platform dedicated to the extraction, transformation and visualization of mobility data. This platform was developed in the framework of a French regional project (DA3T project) aiming at improving the management and valorisation of touristic cities via the fine-grained analysis of tourist mobility data. The system is totally modular, and non-computer scientists (such as geographers) can make processing pipelines from a variety of modules belonging to different categories (e.g., extraction, filtering, visualization, etc.). Each pipeline is created in order to help fulfil one or several reporting needs. Indeed, the results of those pipelines aim to be presented to local authorities and decision makers to assist them in improving infrastructure and tourism management. Beyond this main use case, the platform is also generic and aims to work with any kind of mobility data (not strictly limited to the tourism field). It is heavily inspired by traditional ETL (Extract, Transform, Load) software and processes.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Information systems** → **Geographic information systems**;

KEYWORDS

ETL, geographic information, data integration, mobility tracks, semantic trajectories

ACM Reference Format:
Maxime Masson, Cécile Cayèré, Marie-Noëlle Bessagnet, Christian Sallaberry, Philippe Roose, and Cyril Faucher. 2022. An ETL-like platform for the processing of mobility data. In *The 37th ACM/SIGAPP Symposium on Applied Computing (SAC '22), April 25–29, 2022, Virtual Event, .* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3477314.3507057

SAC '22, April 25–29, 2022, Virtual Event,

ACM ISBN 978-1-4503-8713-2/22/04…$15.00
https://doi.org/10.1145/3477314.3507057

1 INTRODUCTION

Nowadays, advanced data processing and visualization are crucial to analyze and better decipher certain phenomena. Tourism in particular is a field in which the understanding of visitors' behaviors and their type of tourism practices is a particularly important information for urban planners and decision makers at several scales: city, district or even country. In this particular domain, data can be materialized in different ways, for example, the set of trips made during a tourist stay.

The rapid development of mobile technologies and media in recent years has made the collection of mobility data increasingly easy and, above all, in large quantities. But simply collecting data is not enough, they must be processed and then presented to domain specialists who can analyze them. As such, it is particularly important to design and set up frameworks for extracting, transforming and visualizing mobility data.

We therefore propose a modular and generic platform dedicated to the extraction, transformation and visualization of mobility data. This work was done as part of a French regional project (DA3T project) aiming at improving the management and valorization of touristic cities through the fine-grained analysis of visitors' mobility data. This new platform is inspired by traditional ETL (Extract, Transform, Load) tools but is strictly dedicated to the processing of mobility and associated data. The originality of this work lies in the fact that to date, no ETL tool is dedicated to the handling of mobility data which is associated with specific requirements and processes. Moreover, this platform aims to be used by non-specialist users and therefore remain as simple as possible.

The article will be organized as follow. Firstly (**Section 2**), we will introduce the DA3T initiative, the scenario motivating our work (**Section 2.1**) and the requirements for the platform (**Section 2.2**). **Section 3** will introduce the state of the art in ETL tools on which we based our work. Our contribution will then be presented: An ETL-like platform dedicated to the extraction, transformation and visualization of mobility data (**Section 4**). Finally, we will conclude by addressing future perspectives within the framework of the DA3T project (**Section 5**).

2 DA3T PROJECT

DA3T (for *Device for the Analysis of Digital Tracks for the Valuation of Touristic Territories*) is a project aiming at improving the

management, the development and the valuation of tourist cities in the *Nouvelle-Aquitaine* region (France) by the fine-grained analysis of the tourist practices in its cities. It is an interdisciplinary project that integrates both computer scientists and geographers. *La Rochelle* is the support city for some of the experiments carried out in the framework of this project.

First of all, **GPS tracks are collected periodically** using the phone of volunteer tourists through a mobile application called *GéoLuciole* developed as part of the project. The collected data are called **mobility tracks** and provide a comprehensive view of the places visited by these tourists during their stay. We model the movement of tourists in a discrete way via a series of geolocated and time-stamped positions. Each position of a track is represented by a tuple $p = (o, x, y, t, D)$ with p the capture of the moving object o (the visitor), x and y the spatial coordinates (latitude and longitude), t the time stamp of the capture, and D a set of additional collected metadata such as capture accuracy, speed, etc. The user can choose the duration of the data acquisition and stop it at any time. These tracks are discrete due to technical limitations (processing, capture, battery backup, etc.) and vary in precision depending on the type of the capturing device. No personal data is stored in order to comply with the General Data Protection Regulation (GDPR).

The second source of data are **semi-directive interviews** conducted with tourists at the end of their stay. They allow to collect a partial memory of the tourist practice as a qualitative data. The tourists use words (*from, in front, next to, etc.*), and places to describe their practice. These interviews are based on the principle of the *photo elicitation interview* [7], the interview is conducted using the map of the visitor's mobility tracks during his stay so that it feeds the person's speech, this speech then enriches the map itself.

In addition to that, **publicly accessible data from Open Data platforms**, such as DATAtourisme [1] can be used to enrich and complement the tracks (for example: *points of interest, restaurants, weather, etc.*). These contextual data are an important added value for geographers and planners because they help giving meaning to the tourists' movements and therefore facilitate their understanding.

2.1 Motivation scenario

Figure 1 shows a tourist mobility track with 3 different levels of interpretation. Each point represents a timestamped GPS position belonging to the mobility track of a tourist visiting *La Rochelle*.

- The first interpretation (**Figure 1, 1**), is the **raw mobility track** which is a set of spatio-temporal positions.
- The second interpretation (**Figure 1, 2**), is the **mobility track completed with points of interest** (POIs). For example : *Saint-Nicolas Tower, Lantern Tower, etc.*
- The third interpretation (**Figure 1, 3**), is the **mobility track completed with extracts from the semi-directive interviews**. For example: "*we went to the lantern tower, but it was too late*".

[1] https://info.datatourisme.gouv.fr

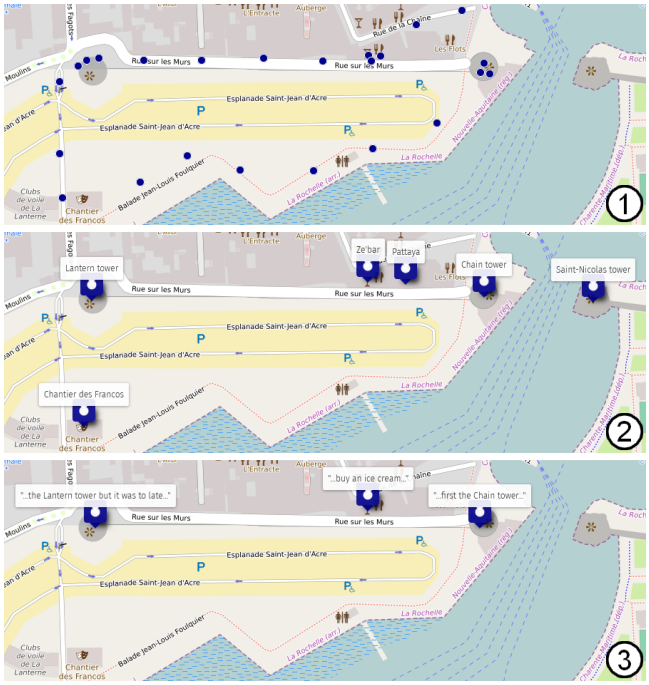

Figure 1: A mobility track with 3 levels of interpretation

In order to help with the handling of these collected data, an important part of the DA3T project is to **design a platform facilitating the exploitation of mobility and associated data (such as interview data and contextual data)**.
The main objective of the platform is to allow non-computer scientists (e.g., geographers, urban planners) to manipulate the data and make analysis on them, so they can report to local authorities regarding the visitors' tourism practices. Below are examples of needs they can have and the different steps to be taken to obtain the desired result.

View all the mobility tracks passing in the *Les Minimes* district and the *St-Nicolas* district.

(1) Get the mobility tracks collected by *GéoLuciole*.
(2) Get the districts of *La Rochelle* (from *La Rochelle* Open Data sources).
(3) Filter the tracks where at least one position is included in the *St-Nicolas* district and one in the *Les Minimes* district.
(4) Display the resulting tracks.

Get all the mobility tracks of 2020 where the weather was always rainy.

(1) Get the mobility tracks collected by *GéoLuciole*.
(2) Filter the tracks captured in 2020.
(3) Get the weather data for 2020 in *La Rochelle* (Open Data).
(4) Filter the tracks of which all the positions' corresponding weather is rainy.

The platform we propose aims to allow them to set up and run this kind of processing pipeline in a simple, user friendly and visual way.

2.2 DA3T platform requirements

In collaboration with the project's geographers who are going to be the main users of the platform, we have identified several requirements for it. It has been decided that initially, the platform will only be able to process spatio-temporal mobility tracks and associated contextual data (we leave aside the data from the semi-directive interviews described in **Section 2**). Firstly, we will address the requirements regarding the processing capacity of the platform (**Section 2.2.1**). Then, the human-computer interaction requirements will be discussed (**Section 2.2.2**).

2.2.1 *ETL Processing modules requirements.*

Extraction. We have 2 main requirements. Firstly, the platform must be able to read and parse the various formats in which the mobility tracks are stored. Secondly, it must be able to access external resources (e.g., Open Data and other API) to retrieve contextual data that can be attached to mobility tracks (such as *points of interest, district, weather, etc.*).

Transformation. For this step, we have defined with geographers a battery of generic processing requirements that can be applied to any mobility track, among those:

(1) **Cleaning**: removing or correcting outlier positions from tracks (they can distort the visualization and analysis of the data).
(2) **Trajectory building**: splitting tracks to make them smaller and so be able to visualize them in a more focused and precise way (those smaller tracks are called *trajectories*).
(3) **Semantic enrichment**: adding context to tracks by attaching external data to them (e.g., districts, points of interest, etc.).
(4) **Advanced filtering**: selecting only some tracks or positions with filters.
(5) **Segmentation**: segmenting a track into episodes (*episodes* are similar to *trajectories* but are built using contextual enrichment data).
(6) **Stop and move detection**: detecting stop and move in tracks.
(7) **Similarity**: evaluating the similarity of several tracks or trajectories.

Loading. In this other step, the platform must offer several advanced visualization modules. We have decided to start by selecting two types: the space-time cube and the visualization of the different enrichment made. Indeed, a fully integrated, all in one tool that does not require external software to analyze the data is needed (the user should have no needs for additional GIS tools or data warehouses).

2.2.2 *Human-computer interactions requirements.*

Below are the requirements defined for the platform regarding human-computer interactions:

- **High level of specification**: dedicated solely to the processing of mobility and associated contextual data.
- **Mobility data genericity**: having the ability to be adapted to any kind of mobility data (not strictly limited to the tourism domain).
- **Ease of use**: easy to use for non-computer scientist users (e.g., geographers and urban planners).
- **Evolutivity**: easy to implement and deploy new modules to users.
- **Flexibility and modularity**: the platform should be flexible and so proposes a modular approach to build processing pipelines (selecting from a palette of modules and assembling them to answer a specific needs).

3 RELATED WORK

For the platform, we inspired ourself from ETL tools. We have chosen in this article to position and compare it with some existing ones. We will therefore not discuss other approaches in the processing of digital mobility tracks but instead will focus on the characteristics of our ETL-like platform, as well as, what makes our work original and innovative in this field.

ETL (**E**xtract, **T**ransform, **L**oad) describes a multi-step process used for data integration, i.e. the transfer of data from multiple heterogeneous sources (file, database, etc.) [17] to a single location (e.g., a data warehouse) according to specific needs. An important part of this process is the transformation of source data present in multiple formats into a single homogeneous format. The goal is to have the final data in a format that is suitable for reporting and analysis.

The **extraction** stage consists of retrieving data from various heterogeneous sources. These sources often have a variety of formats. This data can then be inspected and validated to ensure that the expected value ranges are present. In case of corruption, they can be sent back to the source for correction. Several types of extractions are possible (complete source data, extract of source data, etc.) [17].

The **transformation** step aims at transforming all the different data formats into a unique one that is decided according to the needs. Different types of transformation can be applied to the data: filtering, normalization, duplicate removal, join, aggregation, etc. These transformations are generic and can be parameterized, so ETL makes it possible to avoid hand-coding of transformations

Finally, the **loading** stage is a propagation stage [4]. That is, the data is physically loaded, written to another location, usually a data warehouse. The first load is called the initial load and, in case of additional loads, they can be either incremental or complete (reloading all data).
There are two main deployment types for ETL tools.

- **On-premise deployment**: these ETL must be installed on the user's computer and behave like a regular desktop application.
- **Cloud deployment**: cloud ETL tools are accessible through web applications and run remotely on the cloud.

Numerous software allow to set up ETL processes. We will present some of them from two angles: from the **processing module library** angle (**Section 3.1**) and then focusing on the **human-computer interactions** (**Section 3.2**).

3.1 ETL tools and processing modules

3.1.1 *Business intelligence oriented ETL.*

Business intelligence oriented ETL aim to extract and prepare company data for reporting and business intelligence. Business Intelligence (BI) has many definitions, some more focused on the technological side, others on the managerial side [6] but we can define it as all the techniques and structure for reporting and valuing company data for decision making. Most ETL tools fit in this category (example: *Talend Open Studio, Pentaho*, etc.)

A lot of processing modules are generic and can be found in all BI oriented ETL tools. The extraction modules allow to read most of the file and database formats, to retrieve data through APIs or to read applications. At the transformation level, we can find modules such as the join, the mapping or even the filtering allowing to go from heterogeneous data to a common data format. Finally, for this category of ETL, the loading consists in loading the homogenized data in a data warehouse or a data lake to be able to make reporting. Below is a presentation of some of the BI oriented ETL tools we have been inspired by.

Talend Open Studio [2] is a free and open source ETL tool. The software provides more than 600 components divided into 22 categories. It is based on the Java programming language and makes it possible for users to create their own custom components and share them with other.

RapidMiner [3] describes itself as a data science and machine learning platform integrating an ETL tool. It provides more than 400 modules for analyzing data and allows users to create new ones using the *Python* programming language and submit them to a marketplace [16]. Modules are spread over 8 main categories (ex: *Blending, Cleansing, Modelling*, etc.) and over 35 subcategories. It uses a client-server architecture with the server able to be deployed both on-premise or on cloud infrastructures. It is mainly intended for machine learning, deep learning and can even be used for text mining [8].

Pentaho (also called *Kettle*) [4] is a business intelligence suite released in 2006 which provides both data integration, OLAP services and ETL features [3]. It has nearly 250 processing modules (called *transformation steps*).

3.1.2 Spatial ETL.

Traditional business intelligence oriented ETL tools are inadequate for the processing of geospatial data, **spatial ETL** tools are a specific type of ETL that support them and are dedicated to the extraction, transformation and loading of heterogeneous geospatial data. Common geometry geoprocessing algorithms (such as geometry validation or topology check) are also included. [5]. Some BI oriented ETL also have extensions allowing them to handle geospatial data.

GeoKettle [5] is a metadata-driven spatial ETL tool that support geometry vector data (e.g., lines, polygons, points, etc.) [1] and all the associated processes: centroid, distance, buffer, etc. It aims to bridge the gap between geographic data and business intelligence by helping users to build and update geospatial databases and data warehouses.

FME Destop [6] (for *Feature Manipulation Engine*) is an ETL tool made by SAFE Software specialized in geographic and image data. It offers over 400 transformers spread over 16 categories going from raster processing to spatial analysis.

Spatial Extension for Talend [7] is a plugin which can be installed with *Talend* and allows to transform and integrate data between geographic information systems. It adds support for most of the GIS formats for extraction (PostGIS, Shapefile, KML, etc.) and GIS transformation (buffer, centroïd, area and length, distance, etc.).

3.1.3 Overview.

Name	DA3T Needs	Talend	RapidMiner	Pentaho	GeoKettle	FME Desktop
Type		ETL	Data Science Platform + ETL	ETL	Spatial ETL	Spatial ETL
Extraction (E)						
Common features						
Databases	✗	✓	✓	✓	✓	✓
Common files (CSV, JSON, etc.)	✓	✓	✓	✓	✓	✓
Cloud / Webservice / API (contextual data)	✓	✓	✓	✓	✓	✓
Applications (ex: Twitter, etc)	✗	✓	✓	✓	✓	✓
(Spatial) Geometry file (KML, SHP, GeoJSON, etc.)	✓	✓	✗	✓	✓	✓
Transform (T)						
Common functions						
Mapping	✗	✓	✓	✓	✓	✓
Cleansing	✓	✓	✓	✓	✓	✓
Filtering	✓	✓	✓	✓	✓	✓
Normalization	✗	✓	✓	✓	✓	✓
MapReduce	✗	✓	✓	✓	✓	✓
Join	✗	✓	✓	✓	✓	✓
Stats	✗	✓	✓	✓	✓	✓
Duplicate Removal	✗	✓	✗	✓	✓	✓
(Spatial) Reprojection	✗	✓	✗	✓	✓	✓
(Spatial) Simplification	✗	✓	✗	✓	✓	✓
(Spatial) Measure (length, distance, etc.)	✗	✓	✗	✓	✓	✓
(Spatial) Geometries	✗	✓	✗	✓	✓	✓
Functions needed for DA3T						
Cleaning based on GPS accuracy	✓	✓	✓	✓	✓	✓
Map matching	✓	✗	✗	✗	✗	✗
Spatio-Temporal Trajectory building	✓	✗	✗	✗	✗	✗
Tracks merger	✓	✗	✗	✗	✗	✗
Semantic enrichment	✓	✗	✗	✗	✗	✗
Stop detection	✓	✗	✗	✗	✗	✗
Advanced trajectory filtering	✓	✗	✗	✗	✗	✗
Trajectory similarity	✓	✗	✗	✗	✗	✓
Segmentation	✓	✓	✓	✓	✗	✓
Load (L)						
Common targets						
Data warehouse	✗	✓	✓	✓	✓	✓
Databases	✗	✓	✓	✓	✓	✓
Common files	✓	✓	✓	✓	✓	✓
(Spatial) Geography file	✓	✓	✗	✓	✓	✓
Targets needed for DA3T						
Spatio-temporal cube view	✓	✗	✗	✗	✗	✗
Enrichment view	✓	✗	✗	✗	✗	✗
Map view	✓	✓	✓	✓	✓	✓
Temporal view	✓	✗	✗	✗	✗	✓

Table 1: Comparison of several ETL tool modules with the DA3T needs

Table 1 shows a comparison between 5 ETL tools (*Talend Open Studio, Pentaho, RapidMiner, GeoKettle* and *FME Desktop*) regarding the processing modules offered. The second column correlates these tools with the processing module requirements of the DA3T project platform. A X means that the feature is not present in the software, a ✓ means that the feature is present and an ✓ means that the feature is not present by default but that existing external plugins for the tool allow to add it.

We can see that the 5 ETL tools do not meet all the processing requirements for the project, hence the need to develop a new

[2] https://www.talend.com/products/talend-open-studio/
[3] https://rapidminer.com/
[4] https://www.hitachivantara.com/en-us/products/data-management-analytics/pentaho.html
[5] https://live.osgeo.org/archive/10.5/fr/overview/geokettle_overview.html
[6] https://www.safe.com/fme/fme-desktop/
[7] https://talend-spatial.github.io/

platform. On the one hand, business intelligence oriented ETL tools (e.g., Talend) offer modules that are too generic and not particularly adapted to the sole processing of mobility tracks, on the other hand, spatial-ETL tend to offer overly extensive functionality which are unneeded for our use cases and complicate things for non specialist users. Some of the modules required to process, enrich and view mobility tracks aren't offered by any ETL tool as of now (e.g., space-time cube view, stop and move detection, trajectory similarity, etc.). Last but not least, these ETL tools offer quite a large number of processing modules, most of which have no use for the handling of mobility data, it greatly complicates the use of the software for novice users.

3.2 ETL tools and human-computer interactions

We will now discuss the different types of existing ETL tools from the perspective of human-computer interactions.

3.2.1 *Business intelligence oriented ETL.*

Most ETL tools have a graphical interface structured in the same way and have similar interactions.

(1) A **module palette** allows to browse all the processing modules available. Those modules are splitted into categories (for example: data access, transform, utility, etc.) and sometimes subcategories.
(2) A screen allow to view the **structure of the project** and to add files, or other external resources.
(3) The main area allows the user to build a **processing pipeline**. It is usually based on a node network approach. The user simply drag and drop the modules he wants to add from the module palette screen to this one, and can organize them as he wants. Modules can be linked with connectors. Validity checks are performed to ensure the processing pipeline remains consistent (connectors have to link two compatible processing module parameters).
(4) A **contextual screen** allows to change the parameters of the module currently selected and eventually view it's output.

3.2.2 *Spatial ETL.*

Spatial ETL stand out by offering **screens dedicated to spatial data**. Most of them have a view showing the tree structure of geographic features being processed.

Similarly, the preview tools are adapted to this type of data, instead of showing the results only in tabular form, they often offer a map view with customization options (feature color, toggle feature visibility, etc.)

3.2.3 *Overview.*

To conclude, ETL tools often have a **very wide field of use**, both for enterprise data with business intelligence but also for geographic data in the broad sense. There are even ETL dedicated to text and natural language processing (such as *GATE* [8] or *LinguaStream* [9]). However, to date, there is **no ETL tool strictly focused on the processing of mobility data**.

[8] https://gate.ac.uk/
[9] http://www.linguastream.org/

Name	Talend	RapidMiner	Pentaho	GeoKettle	FME Desktop	DA3T ETL
Type	ETL	ETL	ETL	Spatial ETL	Spatial ETL	Specialized ETL
Common features						
Module Palette	✓	✓	✓	✓	✓	✓
Project Structure Tree	✓	✓	✓	✓	✓	✗
Pipeline Building Area	✓	✓	✓	✓	✓	✓
Tabular Result Screen	✓	✓	✓	✓	✓	✗
Parameter Menu	✓	✓	✓	✓	✓	✓
Spatial features						
Map Preview Screen	✗	✗	✗	✓	✓	✓
Geographic Features View	✗	✗	✗	✓	✓	✗
DA3T needed features						
High level of specification (dedicated solely to mobility and associated data)	✗	✗	✗	✗	✗	✓
Ease of use for non specialist users (few modules, simple UI)	✗	✗	✗	✗	✗	✓
Evolutivity (easy to deploy new modules and updates)	✗	✗	✗	✗	✗	✓
Flexibility and modularity	✓	✓	✓	✓	✓	✓

Table 2: Comparison of several ETL tools features with the DA3T ones

As it can be seen in **Table 2**, this is an issue in our case because, within the framework of the DA3T project, only this type of data has to be processed, and moreover by non-computer scientists users who must therefore have at their disposal a software that is as accessible and clear as possible. This is not the case of the existing ETL tools. They must be as generic as possible and adaptable to any type of data (enterprise data or spatial data). So, they have a rather high complexity of understanding and a lot of modules. It is also quite difficult to make them evolve because it must go through the implementation of plugins that must be redeployed to all users at each change.

4 AN ETL-LIKE PLATFORM FOR MOBILITY DATA

Our contribution is an ETL-like, **modular and generic platform dedicated to the extraction, transformation and visualization of mobility data**. Currently the mobility data supported by the platform are:

- **Spatio-temporal mobility tracks**.
- **Contextual data** (e.g., *points of interest, weather, neighborhoods, etc.*) that can be attached to these tracks using spatial and/or temporal criteria.

Similar to related work, we will present the contribution from the perspective of processing modules (**Section 4.1**) and human-computer interactions (**Section 4.2**).

4.1 DA3T platform processing modules

The platform uses a semantic trajectory model. This model supports both raw mobility tracks but also enriched tracks with multidimensional aspects to them. It is inspired by the MASTER model [13] and is used by the platform for data transfer between processing modules. This model will not be discussed further in this article.

4.1.1 *Processing module categorization.*

We have organized our processing modules into **6 custom categories**. The categories and modules have been decided in consultation with the project's actors (both computer scientists and geographers):

- **Pre-processing**: for modules which aim to extract mobility tracks and prepare them for the future steps.
- **Contextual data**: for modules retrieving contextual data from external API (for example: points of interest of a city, districts, weather, events, etc.).
- **Enrichment**: this category includes all modules related to the enrichment of the raw mobility tracks.
- **Filtering**: for modules used to filter the tracks according to different criteria (temporal, spatial, by property, etc.)
- **Util**: utility modules (for example: merge tracks, limit the number of GPS captures of a tracks, etc).
- **Visualization**: for modules dedicated to visualization.

Table 3 shows a selection of processing modules we have implemented. We can see that most modules required for our platform are highly specialized for mobility data and are therefore not featured in existing ETL tools.

Name	DA3T Category	Module is in existing ETL
Extract (E)		
Mobility Track Extractor	*Pre-Processing*	✗
Context Data Extractor	*Contextual Data*	✗
Transform (T)		
Cleaning	*Pre-Processing*	✓
Map Matching	*Pre-Processing*	✗
Spatial Construct	*Pre-Processing*	✗
Temporal Construct	*Pre-Processing*	✗
Stop detection	*Pre-Processing*	✗
Semantic Enrichment	*Enrichment*	✗
Segmentation	*Enrichment*	✓
Spatial Filtering	*Filtering*	✓
Temporal Filtering	*Filtering*	✓
Property Filtering	*Filtering*	✓
Context Filtering	*Filtering*	✗
Merge Tracks	*Util*	✗
Track Similarity	*Util*	✗
Limit points	*Util*	✗
Load (L)		
Space-time cube	*Visualization*	✗
Map	*Visualization*	✓
Enrichment View	*Visualization*	✗

Table 3: DA3T processing modules and wether or not they exist in other ETL tools

4.1.2 *Extraction modules.*

The **extraction modules (E)** include two types of modules. First, those dedicated to the extraction of mobility tracks. They import the tracks into the processing pipeline by reading databases or files. We have added support for PostgreSQL DBMS and the GeoJSON file format. The second type of extraction modules are those dedicated to the **retrieval of contextual data**. For example: points of interest (via *Google Place*, *DataTourisme* and *Geodatamine* API), districts (via the cities' Open Data platforms), weather (*OpenWeatherMap* API) or even noise zones in cities, etc. These contextual data can then be attached to certain points of the track based on spatial and/or temporal criteria during the enrichment which happens later at the transformation step.

4.1.3 *Transformation modules.*

At the **transformation level (T)**, we have defined a battery of fully generic modules that can be applied to any mobility track imported into the pipeline. First of all, there is the cleaning based on capture accuracy which helps eliminate the positions having a too important radius of uncertainty around them (the threshold is configurable by the user). Another way to clean the tracks is to use the map matching module. It allows to correct outlier positions by attaching the track to the associated network. This network (e.g., *road, pedestrian, cyclists, etc.*) is determined using the movement speed. We have used the *Hidden Markov Map Matching* [14] algorithm for this module.

We then propose modules to build trajectories on spatial criteria (Spatial Construct), for example: *one trajectory per district*. Or even temporal criteria (Temporal Construct), for example: *one trajectory per day, per hour or per week*. Trajectories are subdivisions of tracks that are of interest for a given analysis. Several modules are also dedicated to the detection of stop and move within those trajectories, which are very critical to know when analyzing them but also the similarity between two or more trajectories based on spatial or thematic criteria (such as *Traclus* [12] or *Muitas* [15]).

We also have the semantic enrichment module. The enrichment is the addition of contextual data to raw tracks (e.g., *street names, nearby points of interest, etc.*) through annotations. Contextual data are retrieved using Open Data and other API during the extraction step. They are called *aspects*. Aspects can have a thematic, spatial and/or temporal dimensions, so the semantic enrichment process can attach them to one or several tracks positions (e.g., a position is enriched by a district only if it is included in the polygon encompassing the district, etc.). This process helps to add context to raw tracks and so make them easier to analyze. Lastly, some modules have only an utilitarian purpose, for example merging mobility tracks together (Merge Tracks), limiting the number of positions within a track, etc.

4.1.4 *Loading modules.*

The **loading modules (L)** are strictly limited to visualization. Indeed, as specified in **Section 2.2.1**, we want a fully all-in-one tool that does not require external software for the visualization of the results. Besides the classical map which is universally proposed by spatial ETL tools, we have chosen two visualization modules.

First of all, the space-time cube. It is a 3 dimensional visualization mode aiming to show the behaviors and interactions through space and time. It is based on a cube object in an euclidean space [2]. The width and length (x and y) dimension represent the spatial coordinates (i.e. latitude (y) and longitude (x)) and the height dimension (z): the time. This visualization mode has the advantage of being interactive by allowing the user to change the temporal scale [9]. The tracks (also called *space-time paths*) are displayed within the cube [10]. The second visualization we propose is the Enrichment View, it combines a map and a timeline to display the various enrichment made and on what segment and positions of the track they have been linked to.

Finally, the preview module is an utility module that allows to visualize the result only in the platform without saving it to a persistent file.

4.2 DA3T platform human-computer interactions

As far as the graphical interface and human-computer interactions are concerned, we kept a rather classical structure compared to what is currently done in the ETL field. However, we wanted to simplify it as much as possible in order to make it easily accessible to non-specialist users. Our graphical user interface is divided into 4 main areas (**Figure 3**) which will be detailed here.

4.2.1 *Pipeline list screen.*

This screen, located on the right hand side of the software (**Figure 3, 1**), lists all the processing pipelines created. Each pipeline is associated with additional metadata (creation date, pipeline name, name of the user who created it). The user can create a new one (ADD button), he can also load one from the list or delete it (DELETE button).

4.2.2 *Processing modules palette.*

The palette of available modules (**Figure 3, 2**) is visible at the top of the screen, it works via a tab system. Each tab is associated with one of our module categories and contains all the processing modules of this category. To make everything more visual, we have also associated a unique color to the modules of each category (for example, *the pre-processing modules are associated with the red color, for visualization it is green, etc.*). The user also has the possibility to automatically search for a module using the search bar at the top right of the screen.

4.2.3 *Pipeline building area.*

The main area of the screen (**Figure 3, 3**) is dedicated to pipeline building. It is based on a node network system. The user can drag and drop the modules he wants from the module palette (**Figure 3, 2**). He can then connect these modules together by creating connectors. Built pipelines can be exported to a file using the EXPORT button. The IMPORT is used to import an external pipeline (e.g., load it in the current building area). Finally, the user can also automatically rearrange all the nodes. When a pipeline is being executed once a module is done executing, it reports the number of tracks, trajectories and positions it outputs, as well as it's ending state (success, fail or no output), these information are displayed visually to the user.

4.2.4 *Preview window.*

The preview window (**Figure 3, 4**) is only visible if the user has connected a *preview* module to an output in his processing pipeline. It allows to visualize any output. It can be an HTML file in the case of the space-time cube or the enrichment visualization, but also a text file for the other modules (our semantic trajectory model is implemented with a GeoJSON structure to describe mobility tracks).

4.3 Implementation

We opted for a client-server architecture, we want to be able to quickly update processing modules without having to redeploy the entire platform to all users. **Figure 2** shows the architecture of our platform.

The **server part** (on the left side) contains a Web API (REST), this API exposes all the processing modules which are therefore callable via HTTP requests. The entire modules processing is done directly on the server and the results are returned in HTTP response. The main Web API calls the tracks database for data extraction but can also call external APIs (e.g., *Google Place* [10], *DATAtourisme* [11], *etc.*) to retrieve the contextual data enriching the tracks. In **Figure 2**, module names are represented as numbers to save space.

The **client part** (on the right side) includes a desktop application, it allows the user to build his own processing pipelines, i.e. to choose from the palette of available modules and organize them in the sequence he wants. He can also save his pipelines and execute them. At run time, the client application will make the calls to the modules on the server in a sequential manner. No processing is done locally.

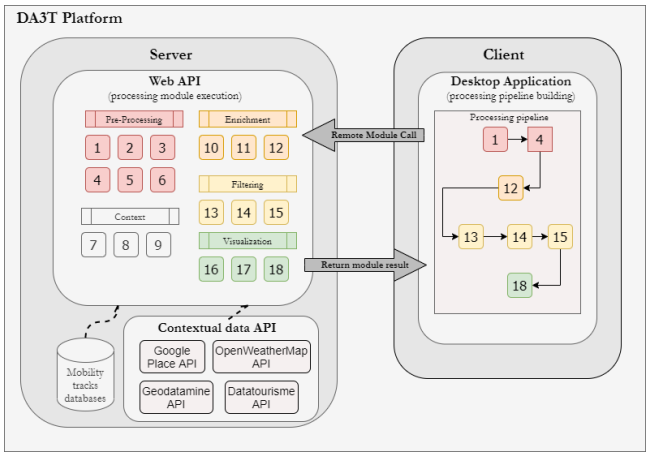

Figure 2: DA3T platform architecture

The server is developed in Python with the Flask web server library. The client is a Windows Presentation Foundation (WPF) desktop application. It has been developed in C# with the .NET Core 3.1 Framework.

4.4 Experimentation

In order to validate the platform and ensure its relevance in the processing of spatio-temporal mobility data, we conducted experiments on several datasets, we will introduce those in **Section 4.4.1**. Following that, we will make a demonstration of how the platform works using an example processing pipeline (**Section 4.4.2**).

4.4.1 *Datasets.*

Firstly, we tested with the *GéoLuciole* data. This is the main dataset of the project, it contains 92 mobility tracks of volunteer tourists collected in *La Rochelle* in 2020 for a total of 120,000 positions, the capture interval is 5 minutes. We will give a detailed processing example in **Section 4.4.2**.

We also tested with mobility data of tourists in New York, these data are from the *FourSquare* social network [18] and represent about 67,000 positions for 193 users; the capture interval is variable, each capture is associated with a particular point of interest (*leisure places, monuments, restaurants, etc.*). This allowed us to observe

[10] https://developers.google.com/maps/documentation/places/web-service/overview
[11] https://info.datatourisme.gouv.fr

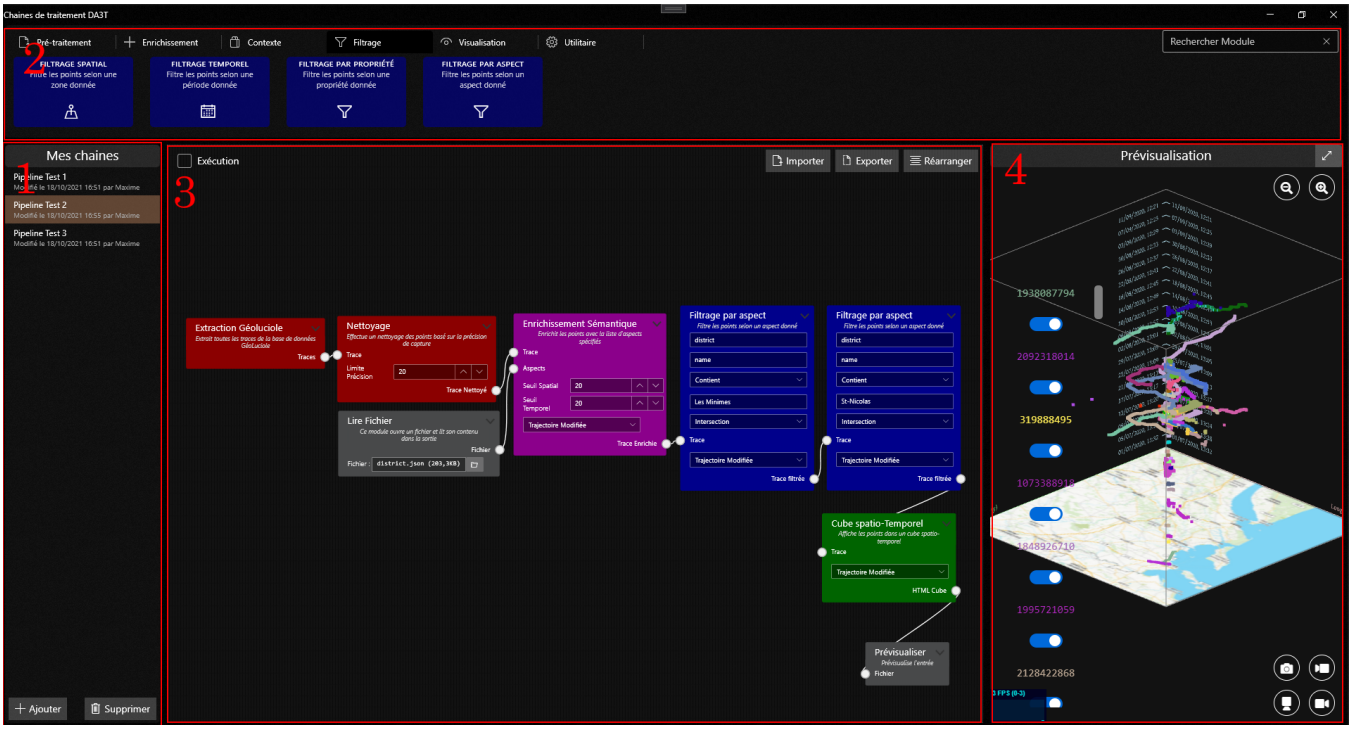

Figure 3: Screenshot of the platform's graphical user interface (GUI)

the dining and leisure habits of a network user over the 14 weeks covered by the dataset.

The 3rd used dataset includes mobility data from 92 pelican migratory birds in the Gulf of Mexico [11] between 2013 and 2016 for approximately 170,000 positions captured at 90-minute intervals. The pelicans are annotated with their colony of belonging, this allowed us to visualize the behavior of pelicans from a particular colony during wintering and late breeding seasons.

Finally, other datasets are planned to be tested with the platform, including runners in town, or movement of naturalists in French Guiana over a large period from 2000 to 2021. We will now present a detailed example of a pipeline made on top of the *GéoLuciole* data.

4.4.2 *Demonstration.*

We will now demonstrate how the platform works, for this we reuse a scenario from **Section 2.1** (*View all the mobility tracks passing in the Les Minimes district and the St-Nicolas district*). We build the processing chain shown in **Figure 4**.

First (1), we extract the *GéoLuciole* dataset. We then perform a cleaning based on the capture accuracy with an uncertainty limit of 20 to eliminate outlier positions which can distort the visualization process. We then enrich the track with the district data (2), each position contained within a district is now annotated with the district information (name and polygon). We also add two filtering modules (3), both are set in *intersect* mode, which means they filter tracks where at least one position meets the required criteria. The first filter keeps only tracks having crossed the *Les Minimes* district, and the 2nd one the *St-Nicolas* district. Finally, we load the result in a space-time cube that we preview (4). There are several other visualization modules available. **Figure 3** display this pipeline after execution in the platform's graphical user interface. The preview screen (**Figure 3, 4**) displays a space-time cube loaded with all tracks crossing both *Les Minimes* and *St-Nicolas* district.

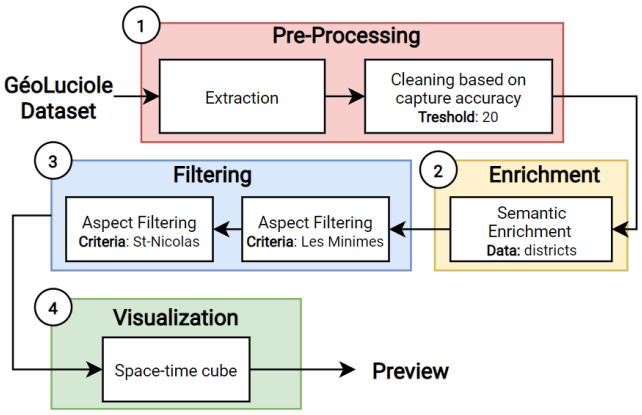

Figure 4: Processing pipeline to get the mobility tracks passing in the *Les Minimes* district and the *St-Nicolas* district

5 CONCLUSION

In this article we presented a **new platform dedicated to the extraction, transformation and visualization of mobility data**.

This platform is inspired from traditional ETL (Extract, Transform, Load) tools and is developed as part of a project (DA3T). The role of this platform is to assist local authorities in making decisions regarding the management of these cities. It is thus highly accessible by non-computer scientists and allows geographers and urban planners to answer needs in analysis and visualization.

As a reminder, this platform is still under active development, one of our main objective is to set up large-scale tests with end-users and a framework to evaluate it on the various requirements defined in **Section 2.2** (e.g., ease of use, etc.). It is also planned to extend this platform so it can **perform text processing and analysis on tourists interview data**. The goal is to provides several modules making it possible to extract specific sentences from tourists' speech and using named entity (places, moment of the day, etc.) recognition to correlate them with segments or positions of the tourist mobility tracks. Lastly, the platform is going to be open source and released to the public in 2022.

REFERENCES

[1] Winda Astriani and Rina Trisminingsih. 2016. Extraction, Transformation, and Loading (ETL) module for hotspot spatial data warehouse using geokettle. *Procedia Environmental Sciences* 33 (2016), 626–634.

[2] Benjamin Bach, Pierre Dragicevic, Daniel Archambault, Christophe Hurter, and Sheelagh Carpendale. 2017. A descriptive framework for temporal data visualizations based on generalized space-time cubes. In *Computer Graphics Forum*, Vol. 36. Wiley Online Library, 36–61.

[3] Roland Bouman and Jos Van Dongen. 2009. Pentaho solutions. *Business Intelligence and Data Warehousing with Pentaho and MYSQL* (2009).

[4] Jaydeep Chakraborty, Aparna Padki, and Srividya K Bansal. 2017. Semantic etl—State-of-the-art and open research challenges. In *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*. IEEE, 413–418.

[5] Urška Drešček, Mojca Kosmatin Fras, Jernej Tekavec, and Anka Lisec. 2020. Spatial ETL for 3D building modelling based on unmanned aerial vehicle data in semi-urban areas. *Remote Sensing* 12, 12 (2020), 1972.

[6] Éric Foley and Manon G Guillemette. 2010. What is business intelligence? *International Journal of Business Intelligence Research (IJBIR)* 1, 4 (2010), 1–28.

[7] Douglas Harper. 2002. Talking about pictures: A case for photo elicitation. *Visual studies* 17, 1 (2002), 13–26.

[8] Markus Hofmann and Ralf Klinkenberg. 2016. *RapidMiner: Data mining use cases and business analytics applications*. CRC Press.

[9] Menno-Jan Kraak. 2003. Geovisualization illustrated. *ISPRS journal of photogrammetry and remote sensing* 57, 5-6 (2003), 390–399.

[10] Menno-Jan Kraak. 2003. The space-time cube revisited from a geovisualization perspective. In *Proc. 21st International Cartographic Conference*. Citeseer, 1988–1996.

[11] Juliet S. Lamb, Yvan G. Satgé, and Patrick G. R. Jodice. 2017. Influence of density-dependent competition on foraging and migratory behavior of a subtropical colonial seabird. *Ecology and Evolution* 7, 16 (2017), 6469–6481. https://doi.org/10.1002/ece3.3216 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/ece3.3216

[12] Jae-Gil Lee, Jiawei Han, and Kyu-Young Whang. 2007. Trajectory Clustering: A Partition-and-Group Framework. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data (SIGMOD '07)*. Association for Computing Machinery, New York, NY, USA, 593–604. https://doi.org/10.1145/1247480.1247546

[13] Ronaldo dos Santos Mello, Vania Bogorny, Luis Otavio Alvares, Luiz Henrique Zambom Santana, Carlos Andres Ferrero, Angelo Augusto Frozza, Geomar Andre Schreiner, and Chiara Renso. 2019. MASTER: A multiple aspect view on trajectories. *Transactions in GIS* 23, 4 (2019), 805–822.

[14] Paul Newson and John Krumm. 2009. Hidden Markov map matching through noise and sparseness. In *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems*. 336–343.

[15] Lucas May Petry, Carlos Andres Ferrero, Luis Otavio Alvares, Chiara Renso, and Vania Bogorny. 2019. Towards semantic-aware multiple-aspect trajectory similarity measuring. *Transactions in GIS* 23, 5 (2019), 960–975.

[16] Petar Ristoski, Christian Bizer, and Heiko Paulheim. 2015. Mining the web of linked data with rapidminer. *Journal of Web Semantics* 35 (2015), 142–151.

[17] J Sreemathy, S Nisha, Gokula Priya RM, et al. 2020. Data integration in ETL using Talend. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*. IEEE, 1444–1448.

[18] Iraklis Varlamis, Christos Sardianos, Vania Bogorny, Luis Otávio Alvares, Jônata Tyska Carvalho, Chiara Renso, Raffaele Perego, and John Violos. 2021. A novel similarity measure for multiple aspect trajectory clustering. In *SAC '21: The 36th ACM/SIGAPP Symposium on Applied Computing, Virtual Event, Republic of Korea, March 22-26, 2021*, Chih-Cheng Hung, Jiman Hong, Alessio Bechini, and Eunjee Song (Eds.). ACM, 551–558. https://doi.org/10.1145/3412841.3441935