International Conference on Automation, Control, Engineering and Computer Science (ACECS'14)
Proceedings - Copyright IPCO-2014, pp.138-147
ISSN 2356-5608

An ETL for Integrating Trajectory Data

A Medical Delegate Activities Use Case Study

Assawer Zekri
Computer science department. BESTMOD Laboratory,
High Institute of Management, University of Tunis,
41, City Bouchoucha, 2000 Bardo, Tunis, Tunisia
assawer.zekri@gmail.com

Jalel Akaichi
Computer science department. BESTMOD Laboratory,
High Institute of Management, University of Tunis,
41, City Bouchoucha, 2000 Bardo, Tunis, Tunisia
jalel.akaichi@isg.rnu.tn

Abstract— **Extract, Transfrom and Load (ETL) represent the core of Business Intelligence (BI). They are responsible for extracting data from heterogeneous sources, transforming and loading them into homogenous targets such as data warehouses. Today, the spatiotemporal and trajectory data warehouses that represent an extension of the classical ones are fed by data coming from the pervasive systems which are equipped with positioning technologies. These latter cause an explosive growth of geo-located data which enable tracking continuously the movements of mobile objects in a time sequence. These data are characterized by their complex structures including spatiotemporal attributes that values represent a data stream generated in a continuous way. In fact, the treatment of the trajectory data which are generated in an unbounded and unpredictable manner, at different speed is difficult. For these reasons, traditional ETL that are based on custom coding for extracting and transforming data become inadequate nowadays. They cannot deal with trajectory data. To solve the associated problems, it is necessary to extend the traditional ETL. So, we propose in this paper a conceptual modeling for trajectory ETL process and trajectory data warehouse. We propose also two algorithms in order to implement trajectory ETL tasks and to construct trajectories. In our work, we use the Geokettle platform to extract geographic points describing the positions of moving objects that are generated from a generator tool. Then, we construct trajectories. Finally, these latter are loaded in a trajectory data warehouse.**

Keywords— Trajectory ETL, Pervasive systems, Trajectory data, Moving objects, Trajectory data warehouse.

I. INTRODUCTION

In order to make the best decision at the right time, the company must adopt the BI process. In fact, this process represents a set of tools and technologies that are used for decision making purposes. The role of these tools and technologies is to collect, store and analyze data in order to extract useful knowledge. So, the collection and the storage of data are done through ETL process. The term ETL stands for extraction, transformation and loading. It is a process that is used to extract data from heterogeneous sources. Then, it cleans and transforms them into an appropriate format. Finally, the cleaned data are stored in data warehouses or data marts. So, in this study, we focus our research on ETL.

Nowadays, the area of trajectory represents a very active research topic. So, the advent of ubiquitous systems and positioning technologies that are increasingly widespread like GPS, RFID and radar allow the real-time monitoring of the movement of mobile objects and generate a sequence of trajectory data that income in unbounded way. The processing of these data arriving in real time, at high speed and in a continuous manner is difficult. In addition, these trajectory data are characterized by their complex structures including spatiotemporal attributes that values represent a data stream to describe the displacement of moving objects over the time. Also, thanks to the huge volume of these data, it is impossible to control the order in which they arrive and to store the stream in whole. This suggests the aggregation structure. Furthermore, the integration of semantic aspect increases the complexity of structure of the trajectory data. In fact, these latter come from various sources (sensors, mobile devices etc). So, the capacity of storage of these sources is limited compared to the huge quantity of generated data. Then, the traditional ETL techniques are not suitable for managing multidimensional data or trajectory data. In addition, the extraction and the transformation of data are done based on custom coding. They consume a lot of time and implementation efforts.

For these reasons, it seems obvious to revise traditional ETL in order to manage the new requirements. This paper focuses on the extension of traditional ETL techniques in order to be able to support the trajectory aspect. So, the goal of our work is an ETL that allows the extraction of geographic points, then, the construction and the aggregation of trajectories to feed a trajectory data warehouse. Our contributions are described as follows:

- The Modeling of trajectory ETL.
- The modeling of trajectory data warehouse that stores aggregated trajectories.
- The validation of our theoretical contribution by integrating the trajectory data in the Geokettle ETL.

International Conference on Automation, Control, Engineering and Computer Science (ACECS'14)
Proceedings - Copyright IPCO-2014, pp.138-147
ISSN 2356-5608

This paper is organized as follows: the section two will present the different research works related to the evolution of data warehouse (Traditional Data Warehouse, Spatial Data Warehouse, Spatiotemporal Data Warehouse and Trajectory Data Warehouse), the evolution of ETL, the conceptual modeling of ETL process and some existing approaches concerning ETL. In section three, we will propose a conceptual modeling for trajectory ETL and trajectory data warehouse based on the case of medical delegate activities. The section four will present two algorithms. The first one concerns the trajectory ETL and the second one focuses on the trajectory construction. The section five will concern the implementation. Finally, we will summarize the work and propose new perspectives.

II. RELATED WORKS AND COMAPARTIVE STUDY

In this section, we will present works which focus on the evolution of ETL, the evolution of data warehouses, the conceptual modeling of ETL process and finally we will present some existing approaches that concern ETL.

A Data warehouse evolutions

According to [1] a data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data to enable decision making. It is a technology that aims to integrate data from heterogeneous sources in a suitable model in order to facilitate the analysis task. Several works exist in this area. So, the previous works such as [2] describe the general architecture of data warehousing. This architecture consists of data sources (bottom layer), an ETL for cleaning and processing data (intermediate layer) and finally data warehouses or data marts for storage (upper layer). Concerning the conceptual modeling of data warehouse, many models are proposed. So, in [3] the author proposes a dimensional model that is composed by a fact table surrounded by dimension tables. Another model proposed by the authors [4] called StarER that combines between the star model and the ER model.

The concept of data warehousing evolves according to technological developments. So, the emergence of ubiquitous systems and positioning technologies constitute a barrier to the traditional data warehouses. Then, it is necessary to extend these classical techniques to support the complex spatial data. These latter appear during the two last decades and refer to geometric objects such as points, lines, polygons, etc. So, in [5] [6] the authors propose conceptual models of spatial data warehouses. In fact, the authors [5] are the first ones to present this concept. They propose a model that supports spatial and non-spatial dimensions and measures. Also, the authors [6] propose a multidimER model for spatial data warehouse. So, several studies exist in this area. But, the spatial aspect is not sufficient for some phenomenon. It must be accompanied with temporal aspect.

Recently, inspired by the rapid development of positioning systems, the spatial data warehouses have been extended to express the spatiotemporal data. But, the spatiotemporal data warehouses are not enough to support the trajectory aspect. This latter requires the semantic or thematic concept to describe the behavior of moving objects. Several works exist in this field. So, we present the proposal of the authors [7]. They describe how the data warehousing technology can be used to store information about trajectories and perform OLAP operations. Furthermore, a model is proposed by [8] that suggests a linear model based on constraints.

After this literature review on data warehouses, we will interest to the ETL in the following section.

B ETL evolutions

As data warehouse, the ETL progresses simultaneously according to technological developments. In fact, it is used to feed a data warehouse.

Many works in the literature focus on the importance of ETL process and describe their functionalities [9] [10] [11] [12]. In fact, the extraction of data is the most important step in a business intelligence project. It allows to connect, read, explore and extract data from heterogeneous sources such as DBMS (Oracle, MySQL, SQL Server, PostgreSQL, etc), ERP (Enterprise Resource Planning), flat files (Excel, XML, Csv) etc. The second step of the ETL process is the transformation. This latter aims to convert and clean data based on rules and operations. The last step of the ETL process is to load the cleaned data in a data warehouse.

According to the authors [13] [14] [15] the first ETL were appeared in the late of 70's. They communicated only with mainframes. In fact, the extraction and transformation of data were done based on custom coding. They were addressed to individual needs. So, in [9][16] authors criticize the traditional ETL. They consumed a lot of time and implementation efforts. Also, the migration of data in a data warehouse, based on traditional ETL, created a latency problem when the volume of data was huge, because they consumed a lot of time and resources. The problems increase with the advent of ubiquitous systems and positioning technologies. These latter generate a massive amount of trajectory data. So, the traditional ETL techniques cannot deal with these data. They treat only simple data and they are designed to feed traditional data warehouses.

Nowadays, the evolution of technologies led to the emergence of new ETL such as spatial ETL, real-time ETL and trajectory ETL. In fact, in Safe Software Incorporation and [17] the authors describe the spatial ETL. It is emerged simultaneously with the spatial data warehouses. It supports the characteristics of geometric objects such as points, lines, polygons, etc. The idea is to make spatial information accessible to users anywhere and anytime. It aims to extract, transform and load spatial data in a spatial data warehouse. But, the spatial ETL does not support the temporal aspect.

To solve this problem, the authors [9] [16] [18][19] refer to the real-time ETL. There can be no real-time data warehouse without real-time ETL. It has to handle update streams. So, the

International Conference on Automation, Control, Engineering and Computer Science (ACECS'14)
Proceedings - Copyright IPCO-2014, pp.138-147
ISSN 2356-5608

collection, transformation and storage of data into a spatiotemporal data warehouse are done in real time in order to have a good quality of data and to make decisions quickly. So, these ETL influence the quality of data. Several works exist in this area, but the spatiotemporal aspect is not enough to support the trajectory phenomenon. So, it must be accompanied with semantic or thematic aspect to describe the behavior of moving objects.

Nowadays, the advent of positioning technologies requires a trajectory ETL to clean data coming from mobile devices etc. They allow the extraction of trajectory data, then transform them by applying rules and cleaning operations. Finally, they are stored in a trajectory data warehouse.

Few studies exist in this area. So, in [20] the author focuses on trajectory ETL which are used to feed the data cube with trajectory data. It proposes two alternatives: a cell-oriented approach and a trajectory-oriented approach. In the first one, the ETL process is described by an algorithm Cell-Oriented - ETL (COE) to feed the fact table of trajectory data warehouse. In this approach, the author seeks the portions of trajectories in all the spatiotemporal cells. Then, the algorithm proceeds to the decomposition of the parties respect to the user profiles. So, the optimality of this alternative depends on the number of portions of trajectories that reside in the spatiotemporal cells. To solve this problem, it proposes a second solution. To avoid checking all the cells, Marketos uses the Minimum Bounding Rectangle (MBR). It is an approximation of trajectories. But, this can results problems when calculating measures because in this case some trajectories can be counted more than once.

To summarize existing works in the area of ETL, we made a comparative study based on some criteria that compares traditional ETL and the new ones as presented in figure 1.

Criterion	Coding		Data Type		The Update		Destination		Extraction		Transformation		Loading		Spatial aspect	Temporal aspect	Trajectory aspect
	Manual	Automatic	Simple	Complex	Immediate	Periodic	Professional	All users	Full	Incremental	Manual	Graphic	Incremental	By reconstruction			
Traditional ETL	x		x			x	x		x		x			x			
Spatial ETL		x		x	x			x		x		x	x		x		
Real time ETL		x		x	x			x		x		x	x			x	
Trajectory ETL		x		x	x			x		x		x	x		x	x	x

Fig. 1. The comparison between the traditional ETL and the new ones

C The conceptual modeling of ETL process

In this section, we talk about works related to conceptual modeling of ETL process. In [21] authors present a conceptual model of ETL based on UML. This approach aims to decompose a complex ETL process into mechanisms such as aggregation, conversion, filtering, joining, merging, etc. The mechanisms are represented by stereotyped classes. The classes are related to each other by UML dependency. An icon is defined for each ETL mechanism. Also, a note is attached to each mechanism to explain its operations and define the mappings between source and target attributes.

According to [22] the conceptual modeling of ETL process can be divided into three approaches: modeling based on mapping expressions and guidelines, modeling based on conceptual constructs and modeling based on UML environment.

Firstly, for the modeling based on mapping expressions, in [23] the authors propose a model based on queries that are used to ensure the warehousing process and present the correspondence between source and target data. So, in this context, the DBMS is considered as a tool of storage and a data transformation engine. Concerning the mapping guidelines, this term refers to the information defined by developers to make the mapping between the attributes of two schemas. Previously, these guidelines are defined manually. Today, they are stored as paper to be useful as references when there is a need to understand how an attribute of a target schema was generated from a source attribute. So, the problem of this modeling is the neglect of graphical model.

Concerning the Modeling based on conceptual constructs, in [24] the authors propose a framework for modeling ETL activities. This framework contains three layers. The first one is called "Schema Layer". It is composed of sources, operations such as control null values and finally data warehouses. All entities of this layer are instances of the five classes: "data type", "Functions", "Elementary Activity", "RecordSet" and "Relationships". These classes constitute the top layer called "Metamodel Layer". The relationship between these two layers is modeled through "instance of". In fact, the "Metamodel Layer" implements the five classes that are necessary to model any ETL scenario. The middle layer is called "Template layer". It is composed of source tables, fact table etc.

Finally, for the modeling based on UML environment, in [25] authors aim to the improvement of UML by modeling the attributes as classes. They offer an approach based on UML package. Indeed, the extraction, transformation and loading of data are modeled as UML packages. In [26] authors propose another model called "Entity Mapping Diagram" (EMD). It supports the integration of different data sources. In fact, the framework (EMD) is composed of three parts: the data sources, the transformation functions (entity, attribute, UDF, conversion of structure) and the data warehouse schema. The entity is a function that can be applied to a source table as eliminating double etc. The attribute can be applied to a source attribute such as "to String" etc. UDF (user define function) is the function that can be added by a user who is the creator of ETL scenario. Finally, the conversion of structure is used to convert the unstructured data sources into structured data.

International Conference on Automation, Control, Engineering and Computer Science (ACECS'14)
Proceedings - Copyright IPCO-2014, pp.138-147
ISSN 2356-5608

D Existing approaches of ETL process

We present an overview of recent approaches existing in the area of ETL. In an article [27] authors propose an approach based on a Multi-Agent System (MAS) which aims to the integration of complex data. The goal of this paper is to exploit the (MAS) which is a smart program composed by a set of agents to perform the task of data integration. So, the extraction phase is done by an agent. It extracts complex data and sends the result to another who is responsible for the structuring step. This step is also performed by another agent who is responsible for the organization of data according to a predefined model. Then, this model is sent to another who is responsible for storage.

In [28] authors contribute to a new approach based on the user profile. This article talks about customizing the ETL phases. In this context, a good understanding of profiles allows companies to know exactly what data to refresh. This approach has several advantages such as saving time, reducing the complexity of tasks etc.

Another approach based on XML. Indeed, the development of this language facilitates the exchange of all types of data. In [29] the author develops an ETL which allows the integration of data in an XML data warehouse. This ETL aims to extract data from heterogeneous sources. Then, two modules are developed. The first one is responsible for establishing a connection with the data sources. The second module aims to convert the data to XML format and loading them in a data warehouse.

The last approach is based on metadata. In fact, complex data require descriptors to define their origin etc. According to [30] an ETL should incorporate a module for managing metadata to define the origins, formats, destination of data etc. Several approaches exist in the field of ETL. Today with the advent of positioning technologies, we need to model a trajectory ETL to clean and manage trajectory data. So, our work focuses on this new area of research.

III. CONCEPTUAL MODELING FOR TRAJECTORY ETL AND TRAJECTORY DATA WAREHOUSE

In this section, we focus on the conceptual modeling for trajectory ETL and trajectory data warehouse.

For the modeling of trajectory ETL, we use UML language. Therefore, in our context, we describe the ETL process as a flow of activities. For this reason, we based on UML activity diagram. This latter provides great expressiveness for modeling. So, our proposed model is composed of two levels. The highest level describes the general process of data warehousing. It consists of three parts. The first one is the extraction of data from mobile devices, then we use the trajectory ETL to clean and transform these data. The last part is loading the cleaned data in a trajectory data warehouse. At the lower level, we detail the trajectory ETL. So, it is considered as a flow of activities. This flow begins with the identification of geographic points (input). Using these points, we construct trajectory portions based on some criteria. These constructed ones are aggregated to regroup the portions and obtain trajectories. Finally, they are stored in a trajectory warehouse. The proposed model is shown in figure 2.

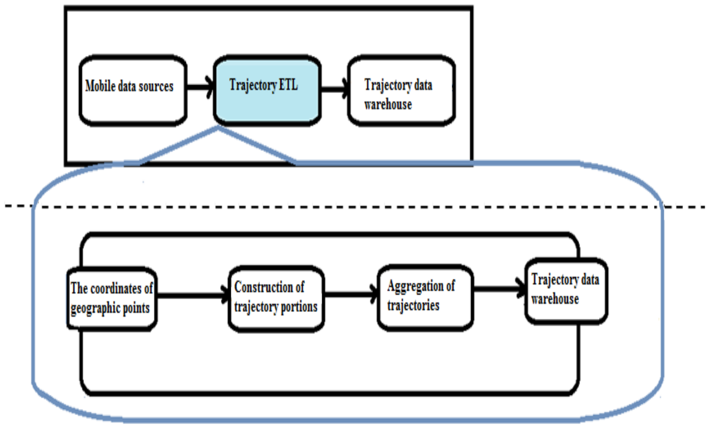


Fig. 2. The proposed model

Next, we propose a generic model that describes the flow of activities of a trajectory ETL process. At the top of the model, we present a significant icon that symbolizes the activity. So, the flow begins with the identification of geographic points (coordinates). Then, these points are stored in a spatiotemporal database. Next, we define the activity to be applied. In our case, the activities are the construction and the aggregation of trajectories. The activity is accompanied by a note that describes its type. Finally, the aggregated trajectories are stored in the trajectory data warehouse (output). The figure 3 summarizes the generic model for trajectory ETL activities.

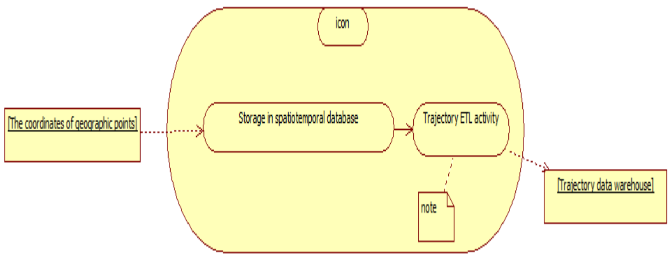


Fig. 3. A generic model for trajectory ETL activity

Then, we propose a conceptual model for construction and aggregation activities. Indeed, the aggregation is one of the main activities of an ETL. In our context, it aims to regroup the trajectory portions that are constructed from the geographic points. Finally, the obtained trajectories are stored in the trajectory data warehouse. The figure 4 presents the conceptual modeling for construction and aggregation activities.

International Conference on Automation, Control, Engineering and Computer Science (ACECS'14)
Proceedings - Copyright IPCO-2014, pp.138-147
ISSN 2356-5608

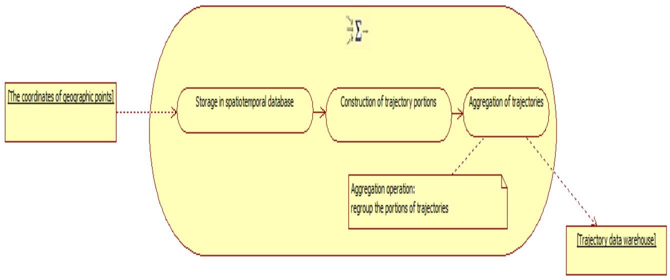


Fig. 4. Conceptual modeling for construction and aggregation activities

The next part of this section focuses on the conceptual modeling for trajectory data warehouse. This latter seems to be a good way for the analysis of trajectory data in a multidimensional context. In our case, we need to analyze the activities of a medical delegate. Firstly, we propose a generic conceptual model that describes the trajectory of mobile object. Then, we model the case of mobile medical delegate. In fact, each one has a positioning device. It uses a means of transport to move. During its movement, it describes a trajectory which is composed by a set of portions. These portions contain moves and stops. The stop may be due to a problem or arrival at destination. The mobile medical delegate belongs to a laboratory that produces drugs. It is responsible for the sale of these products. It moves in one or more road network during a period of time. These roads belong to a region. The UML Class diagram of the trajectory of mobile medical delegate is shown in the figure 5. This description is modeled by classes such as Trajectory, Portion trajectory, Positioning device, Date, Means of Transport, Region, Road network, Destination, Problem, Stop, Move, Mobile Medical Delegate, Laboratory and Drug.

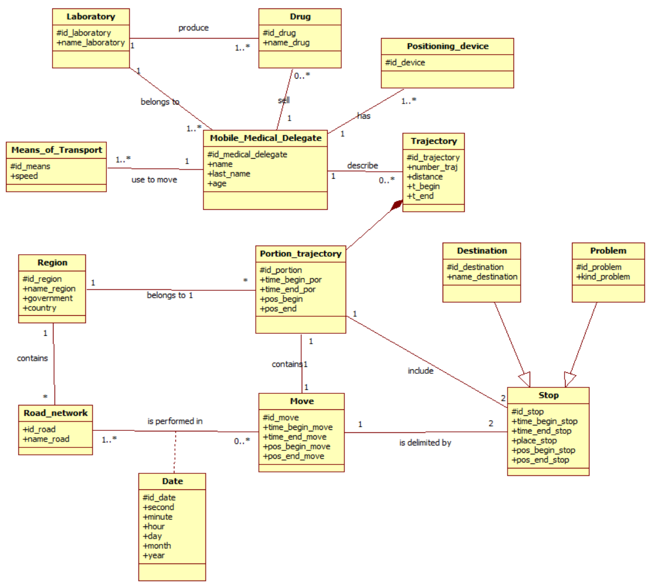


Fig. 5. An UML class diagram of the trajectory of mobile medical delegate

In fact, we choose the snowflake schema to model the trajectory data warehouse. It is a variant and a refinement of the star schema. It keeps the same primitives with normalization of dimension tables to eliminate redundancy. The schema is represented in figure 6.

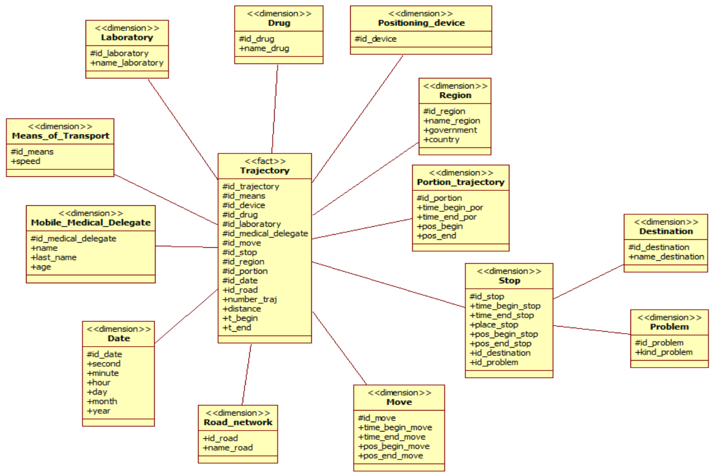


Fig. 6. A snowflake schema of a trajectory data warehouse

For our application we choose to use the snowflake schema for normalization reasons. So, the fact table is:

- Trajectory: the trajectory is the road crossed by a mobile medical delegate. Every trajectory is characterized by several measures such as number of trajectory, distance, t_begin, t_end. It is composed by a set of portions.

Then, we analyze the dimension tables:

- Mobile Medical Delegate: worked in a laboratory and specialized in selling drugs. This dimension includes information about the delegate such as name, last name, age. It is characterized by an identifier (id medical delegate).
- Positioning device: each mobile medical delegate has a positioning device to send data and communicate with the leader of laboratory. It is characterized by an identifier (id device).
- Laboratory: a medical delegate works in a laboratory that produces one or more drugs. This dimension contains information about the laboratory such as name. It is characterized by an identifier (id laboratory).
- Drug: contains information about the drug, such as name. This dimension is characterized by an identifier (id drug).
- Portion trajectory: a trajectory is composed of many portions. Each portion is characterized by a move and two stops. This dimension contains information about the time of beginning for a given portion, the ending time for a given portion, a begin position, an end position. This table is characterized by an identifier (id portion).
- Move: during its course, a mobile medical delegate marks each point of sale. Each move is delimited by two successive stop. This dimension contains information about the time of beginning for a given movement, the ending

International Conference on Automation, Control, Engineering and Computer Science (ACECS'14)
Proceedings - Copyright IPCO-2014, pp.138-147
ISSN 2356-5608

time for a given movement, the begin position of the movement, the end position of the movement. It is characterized by an identifier (id move).

- Stop: it can be detailed with two dimension tables like 'destination' and 'problem. A stop in a trajectory can be the result of a problem, a failure or when the medical delegate arrived at the destination. It contains information about the stop such as the time of beginning for a given stop, the ending time for a given stop, the place of stop, the begin position and the end position of the stop. This dimension is characterized by an identifier (id stop).
- Problem: is defined by an identifier (id problem) and a kind of problem.
- Destination: is characterized by an identifier (id destination) and a name of the destination.
- Means of Transport: is composed of a speed attribute. This dimension is characterized by an identifier (id means).
- Road network: the movement of a mobile medical delegate is done in one or more road network. This dimension is characterized by an identifier (id road) and a name.
- Region: the mobile medical delegate moves in one or more road network that belongs to a region. This dimension is characterized by an identifier (id region), a country, a government and a name.
- Date: This dimension is characterized by an identifier (id date) and attributes such as second, minute, hour, day, month, year.

IV. TRAJECTORY ETL ALGORITHM

New technologies such as GPS, RFID generate streams of trajectory data which arrive in a continuous way. So, it is impossible to store a stream in whole. This requires the aggregation structure.

Then, in our research, we propose two algorithms. The first one implements trajectory ETL tasks and the second one aims to construct trajectories. The trajectory ETL algorithm starts with the generation of positions that describe the movement of mobile objects using a generator tool. These geographic points are stored temporally in a spatiotemporal database. Using the Geokettle platform, we extract points from the table "mobile object", then, we construct trajectory portions. These latter are aggregated and stored in a trajectory data warehouse. These steps are summarized in algorithm 1.

Algorithm 1: Trajectory ETL (geo_point)

Input:

geo_point: spatiotemporal data (geographic points).

Output:

Trajectories stored in the trajectory data warehouse

1:generated_geo_point←Generation_position_moving_objects (geo_point)

2:Table _mobile_ object ← storage_table (generated_geo_point)

3: Point ← Extraction_point (Table_mobile_ object)

4: Portion_Trajectory ← Trajectory_Construction (Point)

5: For all Portion_Trajectory do

6 Aggregated_Trajectory ← Aggregation (Portion_Trajectory)

7: End For

8: Storage (Aggregated_Trajectory, Trajectory data warehouse)

Next, inspired by the work of Marketos, we propose an algorithm for constructing trajectories. This task is difficult. In general, the geographic points are generated in a continuous way, in real time and at high speed. For these reasons, we specify parameters to know if a new point may be added to an existing trajectory or generates a new one. These parameters are: the temporal gap and the spatial gap between two trajectories.

- Temporal gap: is the maximum time allowed between two consecutive time-stamped positions for the same moving object. In fact, any time stamped position of an object received that exceeds the specified temporal gap from its last recorded position will cause a new trajectory.
- Spatial gap: is the maximum distance allowed between two consecutive time-stamped positions for the same moving object. In fact, any time stamped position of an object received that exceeds the specified spatial gap from its last recorded position will cause a new trajectory.

The trajectory construction algorithm aims to build one or more trajectory portions for each moving object to track its movement. This algorithm starts with creating a new trajectory. Then, we add to this trajectory a first Point. Next, any new point p must be compared with the tail of the trajectory and also checked against the specified parameters. If p checks all parameters, then, it is added to the tail of the trajectory. Else, it causes a new one. These steps are summarized in algorithm 2.

Algorithm 2: Trajectory_Construction (Point)

Input :

Point : geographic point.

Output :

Trajectories.

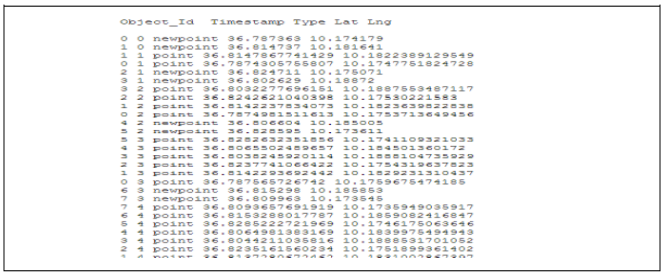
1: For each Id_object ε table mobile_object do
2: Trajectory T ← new Trajectory ()
3 T.add (Point)
4 T.get (Point)

International Conference on Automation, Control, Engineering and Computer Science (ACECS'14)
Proceedings - Copyright IPCO-2014, pp.138-147
ISSN 2356-5608

```
distance ←Calculate_distance (tail, p)
    If distance > spatial gap then
Trajectory T← new Trajectory ()
T.add (p)
   Else T.add (p)
tail ← p
    End If
time ← Calculate_time (tail, p)
    If p.time - tail.time <= temporal gap then
T.add (p)
tail ← p
    Else Trajectory T ← new trajectoire ()
T.add (p)
    End If
End For
```

V. IMPLEMENTATION

Our application is developed in Java using Eclipse, JRE 1.6.0, GeoKettle, PostgeSQL 8.4 with PostGIS 1.4. So, we use a generator tool that generates the positions of moving objects in a text file. In fact, each moving object is characterized by an identifier, type, timestamp, longitude and latitude (the coordinates x and y). The figure 7 describes the generated file.



```
Object_Id  Timestamp Type Lat Lng

0 0 newpoint 36.787363 10.174179
1 0 newpoint 36.814737 10.181641
1 1 point 36.814786774l429 10.1822389129549
0 1 point 36.7874305755807 10.1747751824728
2 1 newpoint 36.824711 10.175071
3 1 newpoint 36.802629 10.18872
3 2 point 36.8032277696151 10.1887553487117
2 2 point 36.824262104039B 10.17530221583
1 2 point 36.8142237834073 10.1823639822838
0 2 point 36.7874981511613 10.1753713649456
4 2 newpoint 36.806604 10.185005
5 2 newpoint 36.828595 10.173611
5 3 point 36.8282632351856 10.1741109321033
4 3 point 36.8065502489657 10.184501360172
3 3 point 36.8038245920114 10.1888104735929
2 3 point 36.8237741066422 10.1754319637823
1 3 point 36.8142293692442 10.1829231310437
0 3 point 36.787565726742 10.1759675474185
6 3 newpoint 36.815298 10.185853
7 3 newpoint 36.809963 10.173545
7 4 point 36.8093657691919 10.1735949035917
6 4 point 36.8153288017787 10.1859082416847
5 4 point 36.8285222721969 10.1746175063646
4 4 point 36.8064981383169 10.1839975494943
3 4 point 36.8044211035816 10.1888531701052
2 4 point 36.8235161560234 10.1751899361402
```

Fig. 7. Generation of positions of moving objects

Then, the generated points are stored temporally in a spatiotemporal database PostgreSQL/PostGIS (Figure 8).

	Object_Id integer	Timestamp integer	Type character(80	Lat [PK] double p	Lng [PK] double p
1	0	0	newpoint	36.787363	10.174179
2	1	0	newpoint	36.814737	10.181641
3	1	1	point	36.8147867741	10.182238912
4	0	1	point	36.7874305755	10.174775182
5	2	1	newpoint	36.824711	10.175071
6	3	1	newpoint	36.802629	10.18872
7	3	2	newpoint	36.8032277696	10.1887553487
8	2	2	point	36.8242621040	10.175302215
9	1	2	point	36.8142237834	10.182363982
10	0	2	point	36.7874981511	10.175371364
11	4	2	newpoint	36.806604	10.185005
12	5	2	newpoint	36.828595	10.173611
13	5	3	point	36.8282632351	10.174110932
14	4	3	point	36.8065502489	10.184501360
15	3	3	point	36.8038245920	10.188810473
16	2	3	point	36.8237741066	10.175431963
17	1	3	point	36.8142293692	10.182923131
18	0	3	point	36.7875657267	10.1759675474
19	6	3	newpoint	36.815298	10.185853
20	7	3	newpoint	36.809963	10.173545

Fig. 8. PostgreSQL database

Next, we connect to the GeoKettle platform in order to extract the stored geographic points (Figure 9).

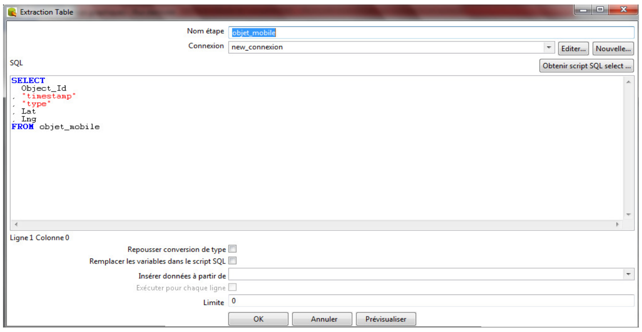


Fig. 9. Extraction points from the "mobile_object" table

After extracting the geographic points, we construct trajectories. The figure 10 shows this.

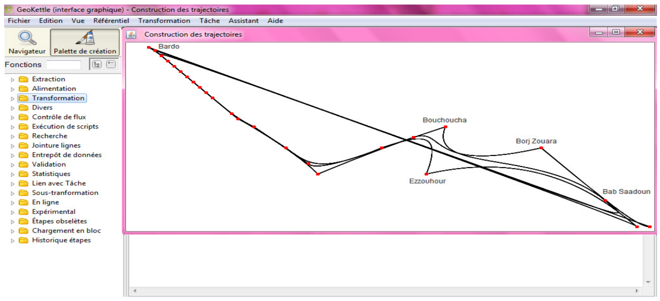


Fig. 10. Trajectory construction

Then, we define our generic flow of activities in the GeoKettle platform as represented in Figure 11.

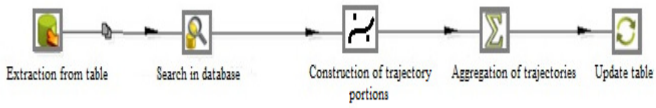


Fig. 11. The flow of activities

The loading data is the last step of an ETL process. After building trajectories, these latter are stored in a trajectory data warehouse. In our application, trajectory data warehouse is composed by a fact table called Trajectory and dimensions such as Mobile Medical Delegate, Drug, Portion trajectory, Date, Destination, Problem, Laboratory, Move, Stop, Means of Transport, Positioning device, Road network and Region. We model our trajectory data warehouse using PostgreSQL. So, the schema is represented in figure 12.

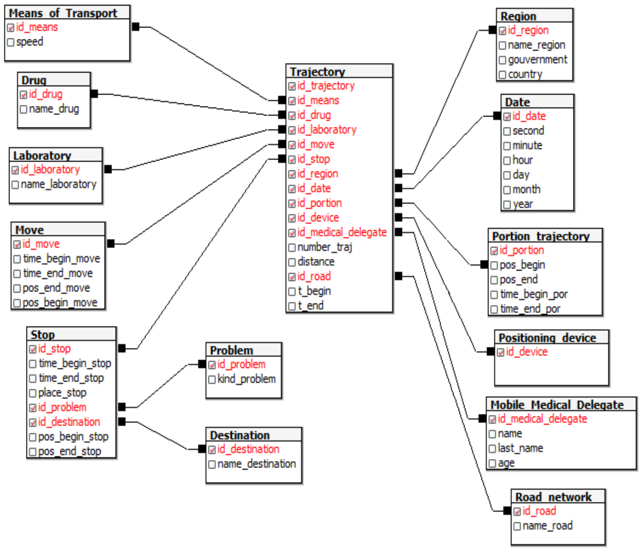


Fig. 12. The snowflake schema of trajectory data warehouse

VI. CONCLUSION

In this paper, we proposed a conceptual modeling for trajectory ETL process and trajectory data warehouse. We also proposed two algorithms in order to implement trajectory ETL tasks and to construct trajectories. Using the Geokettle platform, we extracted the geographic points describing the positions of moving objects generated from a generator tool. The points are stored in a spatiotemporal database to be used for trajectories construction. Finally, these latter are aggregated and loaded in a trajectory data warehouse.

Future work will focus on the analysis and the mining of data stored in the data warehouse in order to extract useful knowledge related to mobile objects activities described by trajectories.

ACKNOWLEDGMENT

We would like to thank everyone, just everyone!

REFERENCES

[1] B. Inmon, “Building the Data Warehouse,” New York/ Wiley, 2nd Ed, 1996, pp. 401.

[2] S. Chaudhuri and U. Dayal, “An overview of data warehousing and OLAP technology,” ACM SIGMOD Record, 1997, pp. 65-74.

[3] J.M. Firestone, “Dimensional Modeling and E-R Modeling In The Data Warehouse,” 1998, pp. 1-9.

[4] F. Busborg, J.G. Borch Christiansen and N. Tryfona, “StarER: a conceptual model for data warehouse design,” In Proceedings of the 2nd ACM international workshop on Data warehousing and OLAP,” Aalborg: University of Denmark, 1999, pp. 3-8.

[5] J. Han, N. Stefanovic and K. Koperski, “Selective Materialization: An Efficient Method for Spatial Data Cube Construction,” In Proceedings of the Second Pacific-Asia Conference on Research and Development in Knowledge Discovery and Data Mining, Simon Fraser: University of Canada, Vol. 1394, pp. 144-158, 1998.

[6] E. Malinowski and E. Zimányi, “Representing spatiality in a conceptual multidimensional model,” In Proceedings of the 12th annual ACM international workshop on Geographic information systems, University of Brussels, 2004, pp. 12-22.

[7] S. Orlando, R. Orsini, A. Raffaet, A. Roncato and C. Silvestri, “Trajectory data warehouses: Design issues and use cases,” In Proceedings of 15th Italian Symposium on Advanced Database Systems, Bari : University of Italy, Vol. 1, pp. 208 – 219, 2007.

[8] F. Giannotti, D. Pedreschi and F. Turini, “Mobility, Data Mining and Privacy the Experience of the GeoPKDD Project,” Privacy, Security, and Trust in KDD, University of Pisa, Italy, Vol. 5456, pp. 25-32, 2009.

[9] K. Kakish and T.A. Kraft, “ETL Evolution for Real-Time Data warehousing,” In Proceedings of the Conference on Information systems Applied Research New Orleans Louisiana, Vol. 5, No. 2214, pp. 1-12, 2012.

[10] T. Oulhaj, “Comment l'informatique décisionnelle peut-elle améliorer la qualité de service client, ” Montpellier, France, 2009.

[11] M. Tranchant, “Capacités des outils SOLAP en termes de requêtes spatiales, temporelles et spatio-temporelle, ” Grenoble, France, 2011.

[12] M. Grimaud, “Comment assurer l'intégration des données structurées dans l'entrepôt de données,” Marketing and Management Department of Information Technology. Higher Institute Of Commerce, Paris, France. 2007.

[13] A.R. Earls, “The State of ETL: Extract, Transform and load technology,” Data informed, 2012. [Online]. Available: http://data-informed.com/the-state-of-etl-extract-transform-and-load-technology/.

[14] F. Morel, “EAI versus ETL:que choisir? ” Journal du net, 2002. [Online]. Available:http://www.journaldunet.com/solutions/0203/020314_eai_vs_etl.shtml

[15]B. Watson, “The Evolution of ETL,” The Third Generation in the Evolution of ETL, 2010. [Online]. Available: http://bvwatson.com/ETL30-blog/2010/07/11/16/.

[16] A. Simitsis and V. Vassilliadis, “Near Real Time ETL,” Annuals of Information systems: New trends in data warehousing and Data Analysis, 2009, pp. 1-31.

[17] K. Zeitouni, “Analyse et extraction de connaissances des bases de données spatiotemporelles. Versailles, France, 2006.

[18] J. Hammer, M. Schneider and T. Sellis, “DataWarehousing at the Crossroads,” Dagstuhl Perspective Workshop, 2004, pp. 1-43.

[19] J. Bernardino and R.J. Santos, “Real-time data warehouse loading methodology,” In Proceedings of the 12th International Database Engineering and Applications Symposium, 2008, pp. 49-58.

[20] G.D. Marketos, “Datawarehousing and Mining Techniques for Moving Object Databases,” Greece, 2009.

[21] S. Luján-Mora and J. Trujillo, “A UML based approach for modeling ETL processes in data warehouses,” Springer, Conceptual Modeling ER, 2003, pp. 307-320.

[22] A.H. El Bastawissy, S.A. El-Sappagh, and A.A. Hendawi, “A proposed model for data warehouse ETL processes,” Computer and Information Sciences, Vol. 23, pp. 91–104, 2011.

[23] N.A. Benharkat and R. Rifaieh, “Query-based data warehousing tool,” In Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP, 2002, pp. 35-42.

[24] P. Georgantas, “A framework for the design of ETL scenarios,” In Proceedings of the 15th international conference on Advanced information systems engineering, 2003, pp. 520-535.

[25] M.S. Lujàn, J. Trujillo and P. Vassiliadis, “Data Mapping Diagrams for Data Warehouse Design with UML,” In Proceedings of 23rd International Conference on Conceptual Modeling, 2004, pp. 191-204.

[26] M. Boshra, A.H. El Bastawesy, A.M.A. Hendawi, “Entity Mapping Diagram as a Conceptual Modeling for ETL Process,” In Proceedings of the third International Conference on Informatics and Systems, 2005, pp. 290-300.

[27] F. Bentayeb, O. Boussaid, O and J. Darmont, “An MAS-Based ETL Approach for Complex data,” In Proceedings of the 10 th ISPE International Conference on Concurrent Engineering, 2003, pp. 49-52.

[28] F. Bentayeb, O. Boussaid, C. Favre, F. Ravat, and O. Teste, “Personnalisation dans les entrepôts de données : bilan et perspectives, ” Actes des 5èmes journées francophones sur les Entrepôts de Données et l'Analyse en ligne, 2009, pp. 7-22.

[29] F.E. Marhoumi, “Entrepôt de données XML: Développement d'un outil Extraction Transformation Load (ETL),” Bruxelles, Belgique, 2006.
[30] L. Demmou, "Exploration de problèmes de performance d'un entrepôt de données, ” Québec, Canada, 2010.