# Data2Help: Data Integration and Cleaning

José Costa
jose.a.costa@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

January 2021

**Abstract**

In Portugal, emergency medical services are coordinated by the Instituto Nacional de Emergências Médicas (INEM). INEM's operational productivity is very important for the country, making the difference in saving citizens' lives. The Data2Help project intends to create tools to optimize the allocation of resources, improving the quality and the response time of medical emergencies. One of the project tasks is the integration of INEM data sources with external data sources. This way, weather data, sports events data, or musical events data can be correlated with the number of occurrences in certain locations. The goal of this task is to develop an integrated data repository, which stores historical data. Consequently, this will allow the findings of data correlations, as well as the application of predictive models using data science algorithms.

**Keywords:** Data Integration, Medical Emergencies, Multidimensional Model, ETL Process, Data Warehouse

## 1. Introduction

In mainland Portugal, emergency medical services are coordinated by the Instituto Nacional de Emergências Médicas (INEM)[1]. Normally, the emergency medical situations are communicated to INEM through telephone calls made to 112, where specialists deal with occurrences and decide what is appropriate for each medical emergency. The Centro de Orientação de Doentes Urgentes (CODU)[2] is the department of INEM responsible for answering 112 calls.

INEM's operational productivity is very important for Portugal, and so it is imperative to reduce the response time to a medical emergency as much as possible, considering that it can save a life. There are two crucial moments in the operations carried out by INEM. The first moment is related to the delay time to answer each call to 112, while the second is related to the time that the emergency vehicle takes from being dispatched until it reaches the emergency.

The Data2Help project aims to create tools to optimize resource allocation, improving both the quality and the response time of medical emergencies in mainland Portugal. The main goals of the project are: (i) forecast the expected workload of CODU; (ii) optimize the schedule of CODU staff to cope with the expected demand; (iii) develop predictive models for the expected demand of emergency vehicles, in different geographic areas; (iv) build software tools to optimize the number of active emergency vehicles and staff across the country, at each work shift.

In order to optimize the operations of emergency medical services, the Data2Help project proposes to develop and apply advanced data analysis algorithms, as well as new models and efficient algorithms, for resource planning and scheduling. In addition to several scientific contributions, a functional prototype will be integrated into INEM in order to test and validate the tools developed in the project.

To achieve the goals, the Data2Help project will integrate INEM data corresponding to occurrences, as well as and other public data, such as weather, sports, and musical events data. This document will focus on developing the task of data integration.

### 1.1. Objectives

The main objective for the data integration task of the Data2Help project is to integrate data sources from INEM with other data sources (*e.g.* sources of weather data, musical events, sports events, among others). As a result, an integrated information system will be built. This system will contain historical data of medical emergencies, response times of medical staff and dispatched vehicles, as well as other information on the operational response of INEM. Moreover, relevant external data that might be correlated with the number of emergency situations in

[1] https://www.inem.pt/
[2] https://www.inem.pt/CODU/

certain locations will also be available.

The Data2Help information system will store and manage historical data that will enable the finding of data correlations, and the learning of predictive models using state-of-the-art data science algorithms. Therefore, it is important to ensure that the information system is capable of answering a well-defined set of queries, developed in straight collaboration with the staff responsible for the next task of the project. Finally, in order to create meaningful prediction models, we must guarantee that the data itself is clean, that is, it does not contain data quality problems such as wrong or incomplete data.

The data integration task will also implement a data refreshment procedure to incorporate new updates, which have been applied to the source databases into the integrated database.

### 1.2. Document Outline

This document is organized in seven additional sections. Section 2 describes the basic concepts for a better comprehension of the main topics: data integration and medical emergencies. Section 3 presents the research work. Section 4 describes the solution developed to perform the data integration. Section 5 presents the experimental validation of the solution. Finally, Section 6 will conclude and present the work that can be performed in the future.

## 2. Basic Concepts

This section presents the basic concepts related to the task of data integration and cleansing in the Data2Help project. These basic concepts will be divided into technical concepts, which in turn are related to data integration (Section 2.1) and domain concepts related to medical emergencies (Section 2.2).

### 2.1. Technical Concepts

Data integration is a set of techniques that allow uniform access to a set of autonomous and heterogeneous data sources, which can be controlled by different people or organizations [4].

The data integration system can implement a materialized data integration approach, aiming to standardize the different data formats present in various data sources.

A materialized data integration system is achieved through the integration of multiple data sources into a single data repository, which is called *Data Warehouse*. This repository aims to store useful information for a given organization, being a part of the decision support process and facilitating the data analysis process [10]. A Data Warehouse should store records from a historical perspective, that is, it must be possible to store data from several years and there must be a time dimension. The Data Warehouse must be refreshed periodically, to load new data from the data sources.

To define the organization of the stored data in a Data Warehouse, the *Multidimensional Modeling* [14] is used. A Multidimensional Model is organized around facts, associated with a set of attributes, which are organized in different dimensions. The *facts* are the focus of what we want to analyze, and they are usually measurable numerical values (*e.g.* the number of units sold in a store or the temperature recorded). The facts can be related to a set of attributes that characterize them from various perspectives. For instance, the temperature can be characterized based on the location and time on which it was recorded. When several attributes are related to the same property of the fact, we have a *dimension*.

The Multidimensional Model can be implemented in a relational database management system. If so, the data is organized into tables, with the facts and dimensions being stored in different tables. The fact tables correspond to the subject we want to analyze, whereas the dimension tables add context to the fact table [10]. There are several types of schemas for this kind of modeling: star, snowflake, or constellation schema.

The *star schema* consists of a single fact table, which forms the center of the star. The fact table forms a *one-to-many* relationship with the dimension tables. Hence, the fact table links to multiple dimension tables, which are in turn linked only to the fact table. The *snowflake schema* is also composed of a fact table linked to multiple dimension tables. It differs from the star schema because it allows a dimension to be represented by more than one table, that is, dimension tables contain normalized data. When a dimension is normalized, some attributes are moved to a new table, which maintains a link to the first. A *constellation schema* is composed of several fact tables that share dimensions.

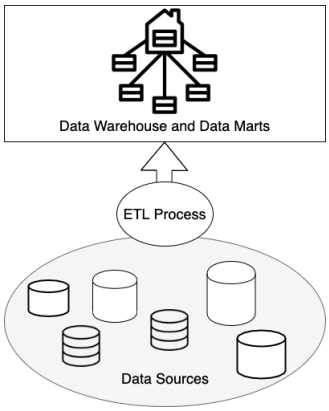


Figure 1: Typical architecture of the materialized data integration process

Figure 1 presents the typical architecture for building a Data Warehouse. The figure shows that the Data Warehouse is the result of running a process on multiple data sources. The process is known as *ETL Process* and it is divided into 3 steps: Extraction, Transformation and Data Loading [4]. In the *Extraction* step the data is extracted from multiple heterogeneous data sources. In the *Transformation* step the data is transformed and modified to ensure that it has the required quality. During this step, data can suffer different types of changes: cleaning, aggregation, formatting, and even new data construction. Finally, in the *Loading* step the data already transformed is loaded to the Data Warehouse. The ETL process is carried out with the support of an appropriated software tool, such as Pentaho Data Integration [3]. As previously mentioned, the end result of the ETL Process is a Data Warehouse that contains information of an entire organization.

Using an ETL process we also can obtain *Data Marts*. Data Marts are smaller Data Warehouses that are specialized in a specific domain or department of an organization. A Data Mart can provide support for decision making in their specific domain.

The main advantage of the materialized data integration and the Data Warehouses is related to the high performance of queries on the materialized basis. However, the main disadvantage is that we need to update the Data Warehouse whenever there are significant changes in data sources [4].

### 2.2. Domain Concepts

In this section, we will introduce some concepts related to Medical Emergencies, to establish a basis of understanding with this domain.

In Portugal, there is the *Sistema Integrado de Emergências Médicas (SIEM)*, which is the set of coordinated actions and entities that cooperate with the aim of intervening and providing assistance to accident victims or sudden illness victims. SIEM is responsible for all emergency activities, such as the prehospital assistance system, transportation, hospital reception, and adequate referral of the patient [7]. The *Instituto Nacional de Emergências Médicas (INEM)* is the entity responsible for coordinating the SIEM.

SIEM is activated when someone calls the European emergency number (112), and the call is answered by the 112 Centers. Whenever the reason for the call is related to a health problem, it is forwarded to the *Centros de Orientação de Doentes Urgentes (CODU)* of INEM.

[3]https://www.hitachivantara.com/en-us/products/data-management-analytics/pentaho-platform/pentaho-data-integration.html

The CODU centers are responsible for quickly answering the call and evaluating the occurrence with the objective of determining the appropriate resources for each occurrence. They also have the responsibility of giving pre-rescue instructions and advice. This whole process is done by doctors and technicians with specific training.

CODU has different means of transport at their disposal, such as ambulances, motorcycles, medical emergency and resuscitation vehicles, medical emergency helicopters, among others. The CODU technicians, through careful analysis, are responsible for activating the different means of transport, supporting them during the provision of assistance to the victims and, according to the clinical information sent by the teams at the occurrence, the CODU technicians are also responsible for selecting and preparing the hospital reception of different patients.

## 3. Related Work

In this section, we will analyze the published work related to both the Data2Help project and the data integration subject. Section 3.1 presents a project developed in London that aims at improving emergency medical services. Section 3.2 presents a platform developed in Spain to support decision-making in emergency situations. Section 3.3 presents lessons learned about the previous projects that can be applied to the Data2Help project.

### 3.1. DASH Project

The aim of the DASH [5] project is to improve the *London Ambulance Service* (LAS). The project explored the potential impact of integrating new data sources and new technologies in medical emergencies. These data sources could be external to the emergency medical services. The team that carried out the project started by analyzing the data that was already stored and used by the emergency medical services. Then, it looked for new data that could be integrated.

The project report contains six new initiatives integrating data to improve the dispatch of LAS ambulances. Suggestions are presented in the following sections:

#### 3.1.1 Health and Social Care data

The DASH project team studied the possibility of integrating health and social care data, in order to bring improvements in the dispatch of LAS vehicles.

LAS didn't collect information about what happens before and after the activation of means of transport in medical emergencies. The DASH project team concluded that the integration of this information could be interesting to assess the LAS approach to previous occurrences and to see what

happens to patients after entering the hospital. This integration could improve the efficiency in the decision-making of the dispatch of ambulances and allowed the development of the project Pre-Hospital Emergency Department Data Sharing (PHED) [3].

Another suggestion from the DASH project is related to the ambulance team's access to patient health records. Thus, there could be improvements in the quality of care and in the decisions of the team at the occurrence, such as a better diagnosis and the referral of patients to an appropriate location for their condition, instead of being referred to the traditional emergency departments of hospitals. An example of the use of health records was the SAFER 2 [12] project, which improved the assessment of elderly people after a fall and if it was appropriated, referred them to the community falls services. In addition to elderly patients, others can be referred, such as patients with mental illness, patients with respiratory problems, or patients with alcohol and drug abuse problems.

The DASH project also proved that it could be interesting for emergency teams to integrate patients' social information. For example, the patients' residence place, possible lack of residence conditions, social care needs, among other information that may be relevant.

#### 3.1.2 Intelligent Transport System data

This initiative aimed to partner with Transport London to provide intelligent mobility for medical emergency vehicles. Transport and smart mobility technologies use traffic data provided in real-time through *crowdsourcing*, via GPS-equipped devices such as smartphones or vehicles. This project considered having a mechanism that recommended the vehicle that would take less time to arrive at the place of the occurrence.

Another proposal that can be followed using smart mobility data is related to the dynamic location of ambulances during a day. A study [11] concluded that the consideration of speeds and travel times in real-time could lead to a repositioning of vehicles during a day, to maintain a reduced response time. The same conclusion was reached in a study developed in Germany [9].

#### 3.1.3 Air Quality Monitoring data

This data integration aimed at an interaction between LAS and the *London Air Quality Network* for resource usage predictions. As is known, air pollution has a significant impact on people's health, especially in urban areas where there is high car traffic. People with respiratory problems, such as asthma, are particularly vulnerable to air pollution and, when they have difficulty breathing, they sometimes need to call for help. Respiratory problems represent a large part of the occurrences, so including improvements in the dispatch for this type of occurrence should be a priority.

The integration of air quality data is quite difficult, but it could bring significant improvements to the services.

#### 3.1.4 Mobile Phone Location data

In this initiative, it was proposed to use data and information from mobile network operators to locate the population in real-time. The use of real-time location data would allow ambulance coverage adjusted to the population's mobility, using spatial analysis and *Data Mining* methods.

The most notorious benefits of integrating this data would be in unforeseen scenarios, when the typical rhythms of mobility in cities were changed, such as unplanned disruptions of transports or natural disasters.

In a study about mobile networks data [2], it was concluded that mobile network operators obtain population mobility data in real-time with a high level of detail. In 2014, a study was developed [1] about the potential use of data from the mobile operator Telefonica[4]. Although the topic of the study was crime, the authors observed that the data collected by the mobile operators provided a spatially-temporal forecast significantly more accurate than historical data.

#### 3.1.5 Video apps

This initiative focuses on the use of data from video communication technology. There are cases in which it is beneficial for the patient to have emergency appointments by video call, since they don't need to go to the hospital, and don't have waiting times there. However, it is necessary to understand if it's clinically safe for the patient. The remote medical appointments will reduce the number of ambulances dispatched, and will consequently allow to have more resources available.

#### 3.1.6 Weather forecast data

This initiative suggests using data from weather forecasts, in order to have a positive impact on the LAS dispatch. The project team looked for trends in occurrences associated with different climatic conditions. These trends were proved in 2014 using data from Birmingham [13], and in 2017 using data from London [6]. It was confirmed that there are more fractures with cold weather and more dizziness and fainting in warmer temperatures. In

[4]https://www.telefonica.com/es/home

the winter it is also natural to have worse road conditions and more staff illnesses, which consequently can cause complications to the services.

### 3.2. Platform to Support Decision Making in Emergency Situations

In a project that took place in Spain [8], an analytical web platform was developed. This platform presents several statistical results of emergencies, allowing users to obtain various information about them, and it was developed for the Canary Islands. The application incorporates occurrence data, as its geographical or temporal location, with data from external sources, such as social and economic data. This allows users to study correlations between external factors and occurrences.

This application allowed improvements in data analysis, helping emergency services to improve their performance in decision-making processes. The platform allows for future integration of new data to provide even more information, which can be related to the occurrences.

#### 3.2.1 ETL Process

The project used historical occurrence data from 2010 to 2014. The most relevant data extracted from the occurrences are: (i) Date and time, with the start and end time of the occurrences; (ii) Age and gender of the people involved; (iii) Location; (iv) Type of the occurrence (*e.g.* road accident, fire); (v) Resources allocated (*e.g.* police, fire); (vi) Assessment of the severity of the occurrence (Low, Medium or High).

It was necessary to carry out an ETL process to develop the platform with the occurrences data. A process for periodic integration of new data was also developed.

#### 3.2.2 Web Platform Development

The next task of the project consisted of building an application to analyze data and improve decision making. The programming language used to develop the platform design was *R* [5], with the support of several libraries. The application allowed an intuitive interaction for users to view geographic data of occurrences on an appealing web page.

### 3.3. Discussion

The DASH project (Section 3.1) has a similar objective to the Data2Help project, both intending to optimize the processes of medical emergency services. In both projects, new data will be integrated to achieve the same objective. For the Data2Help project, some initiatives of the DASH project can be interesting. The integration of health data, social care data, location data, air quality data, or meteorological data in the Data2Help project can allow the detection of correlations between the number of occurrences and these types of data.

The project to develop a data analysis platform to support decision making in emergency situations (Section 3.2) also intends to integrate occurrences data with data from external sources, such as the Data2Help project. Information about the external data sources used and about the ETL process executed in the platform project may be interesting because the information present in both projects is similar.

## 4. Solution

This section describes the data integration solution. Section 4.1 presents the Requirement Analysis task, in which the queries that the integrated data repository must be able to answer were identified. Section 4.2 presents the solution architecture defined after the Requirement Analysis. Section 4.3 presents the intermediate database in which the data is loaded before being loaded into the Data Warehouse. In section 4.4 the multidimensional modeling of the Data Marts that compose the Data Warehouse is detailed. Finally, Section 4.5 describes the ETL process executed to populate the Data Warehouse.

### 4.1. Requirement Analysis

The Requirement Analysis was carried out in partnership with Professor Rui Henriques, who will be responsible for the next task of the project (developing advanced data science algorithms to find data correlations and apply predictive models). This task has the main objective of reaching a consensus between the stakeholders regarding the data integration task.

In this stage, the data to be used for data integration was defined: (i) SIADEM data; (ii) football data; (iii) concerts data; (iv) festivals data; (v) weather data. Then, with the sources defined, a set of queries was identified that should be possible to answer with the integrated data. After the queries were defined, it was concluded that the best method to be used for the data integration would be a materialized data integration, that is, a Data Warehouse would be developed.

### 4.2. Solution Architecture

The solution architecture is presented in Figure 2. As mentioned in the previous section, the data will be stored in a Data Warehouse. The solution starts with an ETL process, which will extract, transform, and upload data from the data sources to the Data Staging Area (presented in detail in Section 4.3). Then, the data goes through another ETL process (presented in detail in Section 4.5), but this time the data is extracted from the Data Staging Area

[5] https://www.r-project.org/

to be loaded into the Data Warehouse. It is important to mention that, before the execution of the second ETL process, it is necessary to develop the multidimensional modeling of the Data Warehouse (presented in detail in Section 4.4).

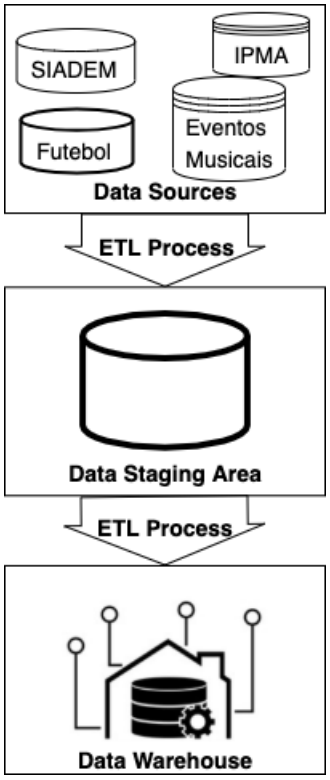


Figure 2: Solution Architecture

### 4.3. Data Staging Area

The data could be extracted, transformed, and loaded directly to the Data Warehouse. However, taking into account the quality and diversity of data sources existing in the Data2Help project, it was defined that the data would be stored in a Data Staging Area, an intermediate and materialized database, in which the data suffers some transformations and validations before being loaded into the Data Warehouse. The external data was extracted, transformed, and loaded using the Pentaho Data Integration and the DBMS used was Microsoft SQL Server. The Data Staging Area is composed of the SIADEM dataset and data extracted from external data sources.

### 4.4. Multidimensional Modeling

With focus on the queries identified in the Requirement Analysis (Section 4.1), a *bus matrix* was developed, which is presented in Figure 3. A *bus matrix* is a tool used to make the Data Warehouse modeling an incremental process, and the lines of the matrix are the business processes. In the Data2Help project, we consider that the business processes are occurrences, that is, facts, and we can insert the occurrences in several subsets depending on the type of information we intend to extract from the Data Warehouse. The columns represent the different dimensions of the Data Warehouse that are shared by the business processes. The *bus matrix* facilitated the association between the business processes and the dimensions. The matrix cells marked with " X " show a logical relationship between rows and columns.

| Dimensions | Time | Location | Emergency type | Unit | Destination | Football | Competition | Teams | Concerts | Festivals | Weather |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Business Processes** | | | | | | | | | | | |
| All occurrences | X | X | X | | | | | | | | |
| Occurrences w/ units | X | X | X | X | X | | | | | | |
| Occurrences w/ complete information | X | X | X | X | X | | | | | | |
| Occurrences / Football | X | X | X | | | X | X | X | | | |
| Occurrences / Concerts | X | X | X | | | | | | X | | |
| Occurrences / Festivals | X | X | X | | | | | | | X | |
| Occurrences / Weather | X | X | X | | | | | | | | X |

Figure 3: Bus Matrix Data2Help

Each business process gave rise to a different Data Mart. Thus, the matrix allowed the development of the Data Warehouse's Multidimensional Model. The Data Warehouse schema is a constellation schema formed by the union of the Data Marts. A constellation schema allows several fact tables that share dimension tables. It is important to note that each entry in the fact tables corresponds to one occurrence.

The dimensions of location, time, and type of emergency are present in all Data Marts. The dimension Location stores the locations used in the Data Warehouse. The dimension Time stores the various time measurements. The dimension Emergency Type stores the type and priority of the occurrence.

#### 4.4.1 Data Mart All Occurrences

The multidimensional model of the Data Mart All Occurrences is represented in Figure 4. The Data Mart includes all occurrences, including those for which no means of transport were activated. The Data Mart is composed of the dimensions Time, Location, and Emergency Type. The fact table has as its primary key the identifier of the occurrence

*id_ occurrence*, while the other attributes of the fact table are the foreign keys of the three dimensions.

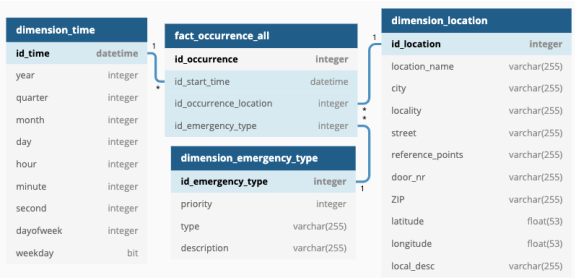


Figure 4: Multidimensional Model of the Data Mart All Occurrences

### 4.4.2 Data Mart Occurrences with Units

The modeling of the Data Mart with the occurrences for which emergency units have been activated, presented in Figure 5, allows extracting information about the occurrences and the units dispatched for these occurrences. In addition to the Time, Location, and Emergency Type dimensions, the Data Mart includes the dimensions: (i) Unit, that contains the list of emergency units; (ii) Group Unit, which is a bridge table between the dimension Unit and the fact table; (iii) Destination, which stores the destination of the first unit dispatched for each occurrence. The primary key of the fact table is the identifier of the occurrence. The other attributes of the fact table are the foreign keys of the dimension tables, the start time of the occurrence, the activation time of the first unit, and the calculation of the time (in seconds) until the activation of the first unit.

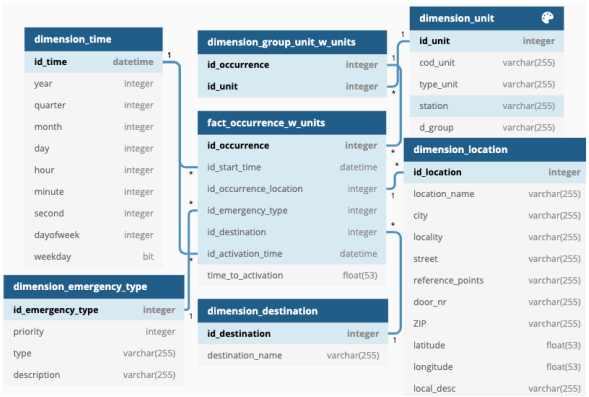


Figure 5: Multidimensional Model of the Data Mart Occurrences with Units

### 4.4.3 Data Mart Occurrences with Complete Information

The Data Mart of the occurrences with complete information is composed of the occurrences for which units have been activated, and for which we have all information about the different times. The dimension tables of this Data Mart are the same of the Data Mart Occurrences with Units (Section 4.4.2). Regarding the fact table, the times of arrival at place of the occurrence, departure from the place of occurrence, and arrival at the place of destination by the first unit are added. Therefore, the times (in seconds) from the activation of the first unit to its arrival at the place of the occurrence, and from the leaving of the place of the occurrence to the arrival at the destination are also calculated.

### 4.4.4 Data Mart Football

The Data Mart Football, represented in Figure 6, is populated by occurrences that may be correlated with football matches. The occurrences loaded are located within a radius of 2 kilometers of stadiums where matches took place in a four-hour period. This period starts 60 minutes before the start of the match. In addition to the dimensions Time, Location, and Emergency Type, there are the following dimensions: Football, that stores data about each match; Competition, which stores the competitions; Teams, which stores the teams. The fact tables are composed by the occurrence identifier (*id_ occurrence*) and the foreign keys of the other dimensions.

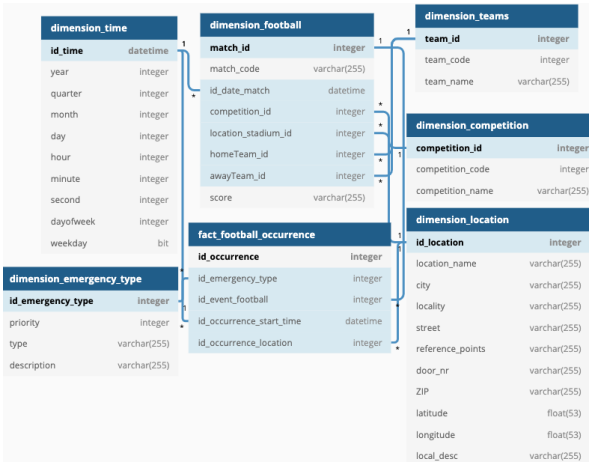


Figure 6: Multidimensional Model of the Data Mart Football

### 4.4.5 Data Mart Concerts

The Data Mart Concerts was modeled to answer queries related to occurrences close to places where

concerts are taking place. It is composed of occurrences located within a radius of 1.5 kilometers from the concert venues, in a six-hour interval, which begins three hours before the scheduled time for the concert's start. In this Data Mart, to the dimensions Time, Location, and Emergency Type, the dimension Concerts is added, which stores data related to each concert. The fact table is composed of the occurrence identifier, as well as the foreign keys of the dimensions.

#### 4.4.6 Data Mart Festivals

In the Data Mart Festivals, occurrences located within a radius of 3 kilometers from the place where the festival took place are loaded. A time interval that starts from the beginning of the day scheduled for the start (12:00 AM), until 11:59 PM of the day scheduled for the end of the festival is used. The Data Mart Festivals is composed of the dimensions Time, Location, Emergency Type, and the dimension Festivals. The difference between the Data Mart Concerts and the Data Mart Festivals is that in the latter, the date of end of the Festival is added.

#### 4.4.7 Data Mart Weather

The Data Mart Weather, presented in Figure 7, is populated by occurrences in areas close to the sensor that made the meteorological measurements, in a time interval of one hour which starts 30 minutes before the measurement, and ends 30 minutes after measurement. The Data Mart is composed of the dimension Weather, which stores meteorological measurements (*e.g.* temperature, humidity, or precipitation), and the Time, Location, and Emergency Type dimensions. The Data Mart Weather fact table is composed of the occurrences identifier and the foreign keys of the dimension tables.

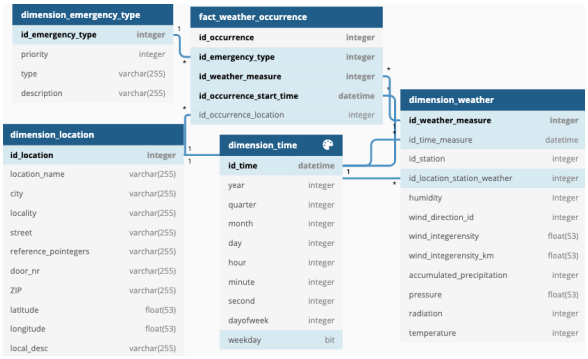


Figure 7: Multidimensional Model of the Data Mart Weather

### 4.5. ETL Process

This section presents the *transformations* of the project's ETL process. The transformations were developed with the Pentaho Data Integration and are composed of *steps* that are executed in parallel and are linked to each other. In the figures along the section, the transformations are represented in a simplified way. Section 4.5.1 presents transformations to populate the Time dimension. Section 4.5.2 presents transformations that load data into the Location dimension. Section 4.5.3 presents the transformation that loads data in the Emergency Type dimension. The following sections present transformations that were necessary to populate other dimensions of the Data Marts. Finally, Section 4.5.6 presents *Jobs* which are processes for orchestrating transformations. Some of the transformations that composed the ETL process are very similar to each other, and as such, only a few will be presented in detail in the document.

#### 4.5.1 Dimension Time

For the dimension Time, all the time measurements that exist in the Data Warehouse were loaded (*e.g.* start time of the occurrences, activation time of the first unit, start time of the festival). The main objective of the transformations that populate the dimension is to fill all attributes of the dimension, through the normalization of the primary key. At an early stage of the transformation, it is necessary to extract data from the Data Staging Area, convert it to *Datetime* format, and homogenize the time zone (in this case to the Coordinated Universal Time (UTC) format). In a second phase, it is necessary to normalize the time measurements to obtain all attributes of the Time dimension. In a last stage, the attributes obtained for the dimension are loaded. The flows of the transformations used to populate the dimension were identical for all time measurements. Figure 8 presents a simplified transformation flow with the steps to insert data into the dimension.

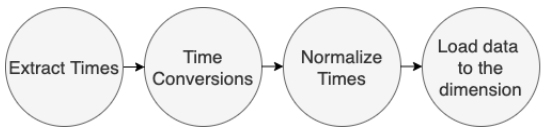


Figure 8: Transformation to populate the dimension Time

#### 4.5.2 Dimension Location

The dimension Location is populated by all the locations of the Data Warehouse (*e.g.* location of the occurrence or location of the football stadium).

The main objective of the transformations to populate the dimension Location is to fill in as many attributes as possible in each table entry, according to the data stored in the original data source. As presented in Figure 9, in the first stage of the transformation, the data is extracted from the Data Staging Area. Then, if required, changes are made in the data to fit the attributes of the dimension. Then, it's necessary to create a *surrogate key* for each different set of latitude and longitude. Finally, the transformation loads the data into the Data Warehouse.

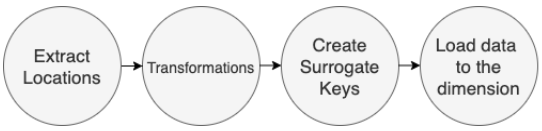


Figure 9: Transformation to populate the dimension Location

#### 4.5.3 Dimension Emergency Type

All occurrences in the Data Warehouse have a type and a priority. The purpose of the Emergency Type dimension, presented in Figure 10, is to contain all existing combinations between the type and priority attributes. For each combination, the dimension will have a different surrogate key. In the first stage of the transformation, the data is extracted from the Data Staging Area. Then, surrogate keys are created. Subsequently, the corresponding description is associated with each attribute type, to simplify the identification of the occurrence. Finally, the data is loaded to the dimension.

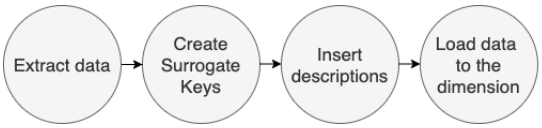


Figure 10: Transformation to populate the dimension Emergency Type

#### 4.5.4 Dimensions Football, Concerts, Festivals and Weather

The transformations used to load data in the dimensions Football, Concerts, Festivals, and Weather are very similar to each other. They start with data extraction from the Data Staging Area. Then, there is the conversion of times to the correct formats. The artificial keys are created for each dimension. After this, it is necessary to look for the artificial keys corresponding to some attributes of the dimension. Finally, the data is loaded into each dimension.

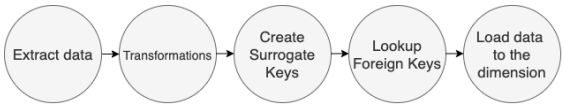


Figure 11: Transformation Model to Populate the dimensions Football, Concerts, Festivals and Weather

#### 4.5.5 Fact Tables

Each Data Mart has a fact table. The model of the transformations executed to populate the fact tables is represented in Figure 12. The transformations begin with data extraction, which can be both occurrences data or external data. Then, occurrences that don't belong to the respective Data Mart are excluded. Finally, the transformation looks for foreign keys that correspond to the remaining attributes of each fact table, and loads the keys into the table.

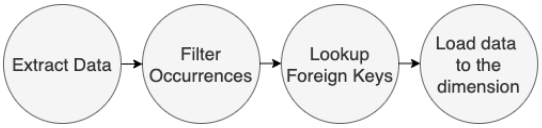


Figure 12: Transformation Model to populate the fact tables

#### 4.5.6 Jobs

In the ETL process of the Data2Help project, *Jobs* were used to sequentially execute the transformations that load the data in each Data Mart. Hence, there is one job per Data Mart. The *Jobs*, as well as the transformations, are composed of steps that are linked to each other and can be executed several times to refresh the Data Warehouse. The execution of *Jobs* can be scheduled to start automatically and thus keep the Data Warehouse updated. Regarding the sequential order of execution of transformations in a *Job*, it is important to note that dimensions without foreign keys must be executed first. Moreover, data can only be loaded for one dimension with a foreign key from another dimension after the data has been loaded for the second dimension. The example of the *Job* model used to load data into the Data Mart All Occurrences is presented in Figure 13.

## 5. Experimental Validation

This section presents the experimental validation of the Solution described in Section 4.

To validate the Data Warehouse, it is necessary to verify that it can answer the queries identified in


Figure 13: Job Model to execute the transformations to populate the Data Mart All Occurrences

the Requirement Analysis (Section 4.1). This validation was successful, considering the Data Warehouse of the Data2Help project is able to answer all identified queries, some through a traditional *SQL* query, others by Pentaho Data Integration transformations used to extract and present the results.

Another necessary validation for the Data Warehouse is related to the performance of the queries executed on the Data Warehouse, in comparison with the equivalent queries executed on the SIADEM database. To carry out this validation, the identified queries that could be executed on the two databases were executed in SQL. Then, a comparison between the processing time of the queries in each of the databases was made. The results proved that the processing time of the queries made on the Data Warehouse is clearly shorter. This happens for several reasons. For example, the Data Warehouse is prepared for aggregations, and because to execute the queries in the SIADEM database, some conversions are necessary.

The third necessary validation is related to the data loaded into the Data Warehouse. It needed to be verified that there was no data loss or incorrect data. For this validation, a Pentaho Data Integration Job was executed. The Job compared the results of queries performed on the two databases and concluded that the results were the same.

## 6. Conclusions

As mentioned in the introduction (Section 1), the productivity of INEM's operations is very important for Portugal, so any improvements in productivity can save human lives. The goal of the Data2Help project is to improve the performance of INEM's processes.

This document presented the task of Data Integration and Cleaning in the Data2Help project. The main objective of the data integration task was to integrate historical data from SIADEM with external data sources, in a repository that would allow answering a defined set of queries. To accomplish the task, a materialized data integration was carried out, which gave rise to the Data Warehouse. In the next stage of the project, predictive models will be applied. Based on the predictive models and the correlations detected in data, new models and algorithms will be developed in order to optimize the planning schedules for both INEM staff and emergency vehicle location.

To conclude, it is possible to state that the main objectives of the data integration task have been successfully achieved. The Data Warehouse allows answering the queries identified in collaboration with those responsible for the next task of the project. Regarding the data loaded in the project, it is also possible to verify that there were no data losses, along with some transformations that were carried out to improve the quality of the data, as well as a mechanism implemented to allow periodic updating of the data. In turn, concerning the performance of the consultations executed on the Data Warehouse and the relational database of SIADEM, it is also possible to conclude that significant performance improvements have been achieved.

### 6.1. Future Work

The data integration carried out still has some aspects that can be improved, especially regarding external data. Some external data such as musical events data or meteorological data are not extracted automatically, and so only a limited set of these data is used. It would be important to use APIs that allow adding new data automatically. The meteorological data used in the project also has the limitation of only corresponding to sensors located in Lisbon.

Still on the topic of external data, it would be interesting to add other data sources, such as data about epidemics. For example, integrate data about *COVID 19* [6]. Unfortunately, the SIADEM data set was inserted in the Data Warehouse prior to the registration of the first person infected with the *SARS-CoV2* virus in Portugal. Certainly, the pandemic data would bring very interesting information to correlate with the occurrences. Still on epidemic data, integrating reliable data about the common flu could also bring relevant information.

Another interesting proposal would be the development of a user-friendly application with an intuitive interface that allows to analyze and have a graphic visualization of the data.

**Acknowledgements**

I'm very grateful to my thesis advisors Professor Helena Galhardas and Professor Vasco Manquinho for the support during this difficult year of 2020. I would also like to thank my family, friends, and girlfriend, for all the support over these years.

[6] www.who.int/covid-19

**References**


[1] A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi, and A. Pentland. Once upon a crime: towards crime prediction from demographics and mobile data. In *International Conference on Multimodal Interaction*, pages 427–434. ACM, 2014.

[2] F. Calabrese, L. Ferrari, and V. D. Blondel. Urban sensing using mobile phone network data: a survey of research. *Acm computing surveys (csur)*, 47(2):25, 2015.

[3] S. Clark, M. Damiani, H. Dorning, M. Halter, and A. Porter. Patient-level data linkage across ambulance services and acute trusts: assessing the potential for improving patient care. *International Journal of Population Data Science*, 1(1), 2017.

[4] A. Doan, A. Halevy, and Z. Ives. *Principles of data integration*. Elsevier, 2012.

[5] A. Drake, A. Pollitt, E. Sklar, L. Smith, S. Parsons, and E. Schneider. Data for ambulance dispatch. Technical report, Policy Institute at King's College London, 2018.

[6] M. Mahmood, J. Thornes, F. Pope, P. Fisher, and S. Vardoulakis. Impact of air temperature on london ambulance call-out incidents and response times. *Climate*, 5(3):61, 2017.

[7] A. P. G. Martins. Emergência pré-hospitalar. 2011.

[8] C. J. Pérez-González, M. Colebrook, J. L. Roda-García, and C. B. Rosa-Remedios. Developing a data analytics platform to support decision making in emergency and security management. *Expert Systems with Applications*, 120:167–184, 2019.

[9] M. Reuter and W. Michalk. Towards the dynamic relocation of ambulances in germany: The risk of being too late. In *2012 Annual SRII Global Conference, San Jose, CA, USA, July 24-27, 2012*, pages 642–649, 2012.

[10] M. Y. Santos and R. Isabel. *Business Intelligence da informação ao conhecimento*. Lisboa: FCA, 2017.

[11] V. Schmid and K. F. Doerner. Ambulance location and relocation problems with time-dependent travel times. *European Journal of Operational Research*, 207(3):1293–1303, 2010.

[12] H. A. Snooks, R. Anthony, R. Chatters, J. Dale, R. Fothergill, S. Gaze, M. Halter, I. Humphreys, M. Koniotou, P. Logan, et al. Support and assessment for fall emergency referrals (safer) 2: a cluster randomised trial and systematic review of clinical effectiveness and cost-effectiveness of new protocols for emergency ambulance paramedics to assess older people following a fall with referral to community-based care when appropriate. *Health technology assessment*, 21(13), 2017.

[13] J. E. Thornes, P. A. Fisher, T. Rayment-Bishop, and C. Smith. Ambulance call-outs and response times in birmingham and the impact of extreme weather and climate change. *Emerg Med J*, 31(3):220–228, 2014.

[14] A. Vaisman and E. Zimányi. *Data warehouse systems*. Springer, 2014.