# DATAMART FOR THE CASE OF EMERGENCY MANAGEMENT AGENCY (EMA)

**Zijadin Krasniqi, PhD (C)**
Informatics, Mathematics, and Statistics
European University of Tirana, **ALBANIA**

## ABSTRACT

Since we do not have a centralized database for data management within the Emergencies Management Agency (EMA), there is a great delay in the exact data entry in real time. From all these departments within EMA, we have difficulties in quick and accurate notification regarding their generated data. This necessarily represents the need for data integration from all these sources, creating thus a unique communication system for EMA itself. Aside from data integration and building the DWH system, it is also needed the actual databases modeling which are built or must be built in EMA's own departments, this can be achieved by studying the DataMart model, and all the processes that will make possible the information gathering from the original sources, and the transformation and upload to the destination. So EMA can facilitate the decision making process in situations when it is directly involved in managing natural disasters, and such situations must often be continuously monitored in order to provide warnings and to take adequate measures for the management of natural disasters. Today, more than ever there is a great need to create a coordinated institutional system with all the resources and capabilities we have available to be prepared and to react in coordinated and planned ways in case of natural disasters and other incidents.

**Keywords:** Data Warehouse, Data Modeling, DataMart, Natural disaster management.

## INTRODUCTION

In order to manage emergencies and natural disasters, it is necessary for the agencies to gather and analyze the massive data of natural disasters. Detailed records of earthquakes, floods or storms and extreme temperatures are the hottest topic in exploring the insight about emergencies and natural disasters management. The analysis that must be generated from them are very diversified. For this reason, the DataMart modeling for this data is a very interesting topic. In order to be built, it is important to use a model which is very easy to understand even from a simple user, because these users will play an active role in generating results. We include here all the steps, in details, which we will follow to build a DataMart based on these data. There are mentioned the main articles and problems which must be solved during the building of models. In order to gain more flexibility regarding the analysis on the saved information, we use the OLAP means of BI technology. (Juliana  2014)

## LITERATURE REVIEW

This part of the study represents the theoretical side of building the DWH which is very welcomed for the decision making processes in business, telecommunication and health care fields. But it also has a great impact in natural disasters management, and this actually needs real time data. DWH systems, from the developing point of view, are very evolutionary and dynamic, they never stop transforming following the requirements. The DWH system is a system that gathers information from different heterogeneous sources. (Kimball 2008)

## ANALYSIS OF DATA FROM KOSOVO HEALTH INSTITUTION - KHI

The health informative system is mainly manual, on paper and most patient’s data and treatments are recorded in paper forms which are gathered in specific institutions and then it is entered in electronic format from operators in hospitals and main centers of family health care. Personnel patients data (the so-called "midnight statistics"), are mainly entered in Excel. However, since 2002, data for activities, possibility of sicknesses and deaths in the primary, secondary and tertiary care are increasingly being entered in the main database "Access-Master Database" Until now, the Access-Master Database is implemented in all QKMFs, in 7 hospitals (5 regional hospitals and 2 city hospitals) and in QKUK. However, data is reported with six months or more delay, the analysis of the same data give different results and not all health institutions are reporting. (Google 2010)

## ANALYSIS FROM KOSOVO SEISMOLOGIC INSTITUTION - KSI

Most of the data from Kosovo Seismic Institution - KSI, are archived in forms of reports, tables, drafts and maps. Their utilization is done with traditional methods. Means of communication are manual, there is no continuous and updated flow gathering, assessment and administration. It is impossible for users to provide new and reliable information in real time about the issues in the seismic field where they are interested in.

## DATA ANALYSIS IN THE KOSOVO HYDRO-METEOROLOGICAL INSTITUTION –KHMI

Rainfall, snowfall levels in high mountains, river flows, underground water levels and levels of accumulative potentials of lakes are all included in this data. All this variables must be continuously monitored and reported from the center to OCEMA which is responsible for the national monitoring of floods. (Google.2011). The quality and reliability of the data are very important. Incorrect or incomplete rainfall data are more harmful than the absence of data. Nothing can be predicted, planned or managed without the data. From the Kosovo Hydro-Meteorological Institution until now we do not have a database which we can refer to in real time and have access to data updates from the Hydro-Meteorological Institution.

## SELECTED DATA MODELING

The analysis of the above requirements leads the data modeling process to build a solution. The very dimensional modeling was selected to be applied in our case, for three main reasons:

- These models offer usage simplicity even from persons which have no knowledge in IT.
- It satisfies complex requirements with high performance.
- It adapts better with high volume of information, as it is in our case. (Juliana 2014)

The DataMart’s that we will be using now serve for performance optimization, for the well-defined and predictable uses, sometimes even with one or without requests. For example, the case of emergencies management could have a DataMart for *Kosovo Health Institution –KHI*, a DataMart for *Kosovo Seismologic Institution - KSI*, a DataMart for *Kosovo Hydro-Meteorological Institution - KHMI*, and so on in order to help the analytic process of reports (Jeffrey 2011).

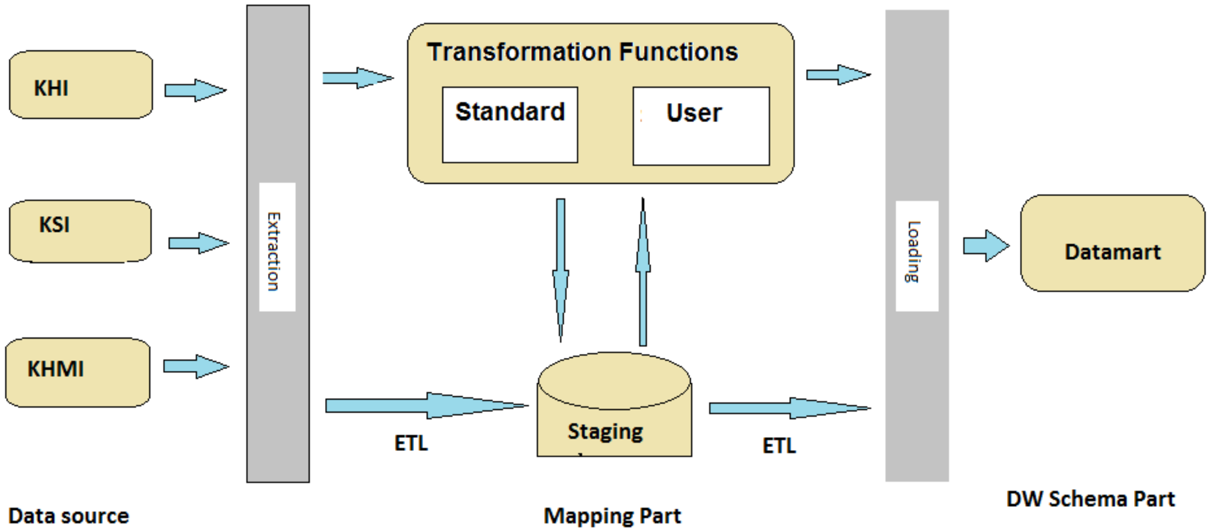
They are built in the same way for other types of records. Also, taking into consideration the problem description and by comparing the elements of the two main methodologies in building the solution, it was selected the Kimball methodology which is based on DataMart's. The criterion elements of selection are represented in the table below:

Table: The criterion elements for modeling selection (Juliana 2014)

| Elements | Kimball | Inmon |
|---|---|---|
| Need | Immediate | Extended in time |
| Focus | Specific business field | All enterprise |
| Budget | Small budget | Big budget |
| Users | Specific business groups | Corporation |

Based exactly on Kimball suggestions, it was selected the double layer architecture, to accomplish especially the partition feature, where data will initially be uploaded to a Stage zone, where it will be transformed, in order to be uploaded later in the DataMart scheme. (Juliana 2014)

Fig: The scheme of selected architecture (Shaker 2011).



We later analyze one of the key techniques which the data structures within DWH are built and implemented with, and this has to do with dimensional modeling.

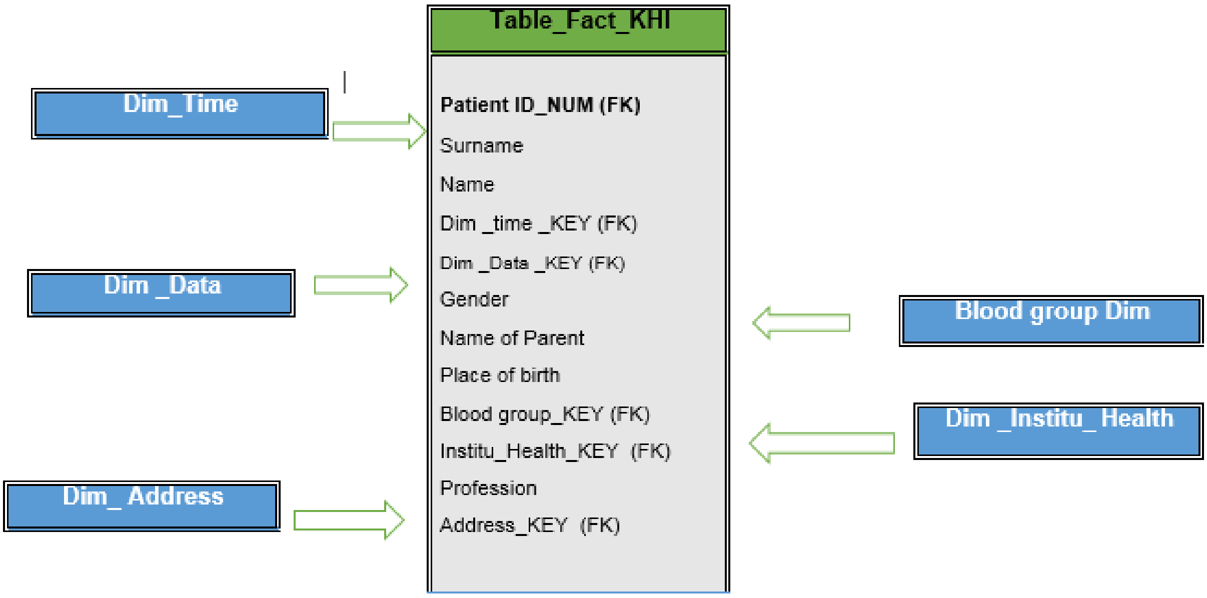
## DIMENSIONAL MODELING

Is a logical design technique for structuring data so that it's intuitive to business users and delivers fast query performance. Dimensional modeling is widely accepted as the preferred approach for DW/BI presentation. Practitioners and pundits alike have recognized that the data presented to the business intelligence tools must be grounded in simplicity to stand any chance of success. Measurements are usually numeric values; we refer to them as *facts*. (Kimball 2008).

The table of facts is composed of main keys with many parts, where numeric and additional facts are the most useful in the table of facts. Each table of facts shows a number of smaller tables called dimensions, the main key of which belongs exactly one of the keys with many parts in the table of facts. Such a description is called the star scheme model, which initially serves for a DataMart. While DWH grows, there can be added more DataMart's and this allows to build more dimensions which are common for more than one DataMart. These common dimensions (CD) act as 'data bridges' between tables and facts, allowing requirements which act in a way that data from multiple facts can be written. Typical examples of common dimensions include date and location.

Therefore, a very important step in building the multidimensional model is the creation of a conformed dimension so that DWH can function as single deposit (Arwa 2011). Each of the processes of emergencies management, as in the case of Kosovo Health Institution –KHI can be shown from a dimensional model which is composed of a table of facts which includes the numerical measurements surrounded by a series of dimensional tables which contain the textual context, as illustrated in the picture about the processes of dimensional modeling

Fig: The processes of dimensional modeling



This structure is very similar to the star, a term dated from the earlier days of relational databases. The dimensional models saved in the relational database platform are generally called "*The star scheme*", while the dimensional models saved in the structures of multidimensional online analytic process are called *cube.*

## FACT TABLES

Fact tables store the performance measurements generated by the organization's business activities or events. The term fact refers to each performance measure. You typically don't know the value of a fact in advance because it's variable; the fact's valuation occurs at the time of the measurement event, such as when an order is received, a shipment is sent, or a service problem is logged.

Nearly every fact is numeric. The most useful facts are both numeric and additive. Additivity is important because BI applications seldom retrieve a single fact table row; queries typically select hundreds or thousands of fact rows at a time, and the only useful thing to do with so many rows is to add them up. Fact tables are huge, with millions or billions of rows, but they're efficient. Because fact tables often store more than 90 percent of the data in a dimensional model, we're careful to minimize redundant data in a fact table. Also, we strive to store the measurements resulting from a business process in a single fact table that's shared across the organization. Because measurement data is the most voluminous data, we purposely avoid duplicating the detailed metrics in multiple fact tables around the enterprise, as we further describe with the enterprise data warehouse bus architecture.

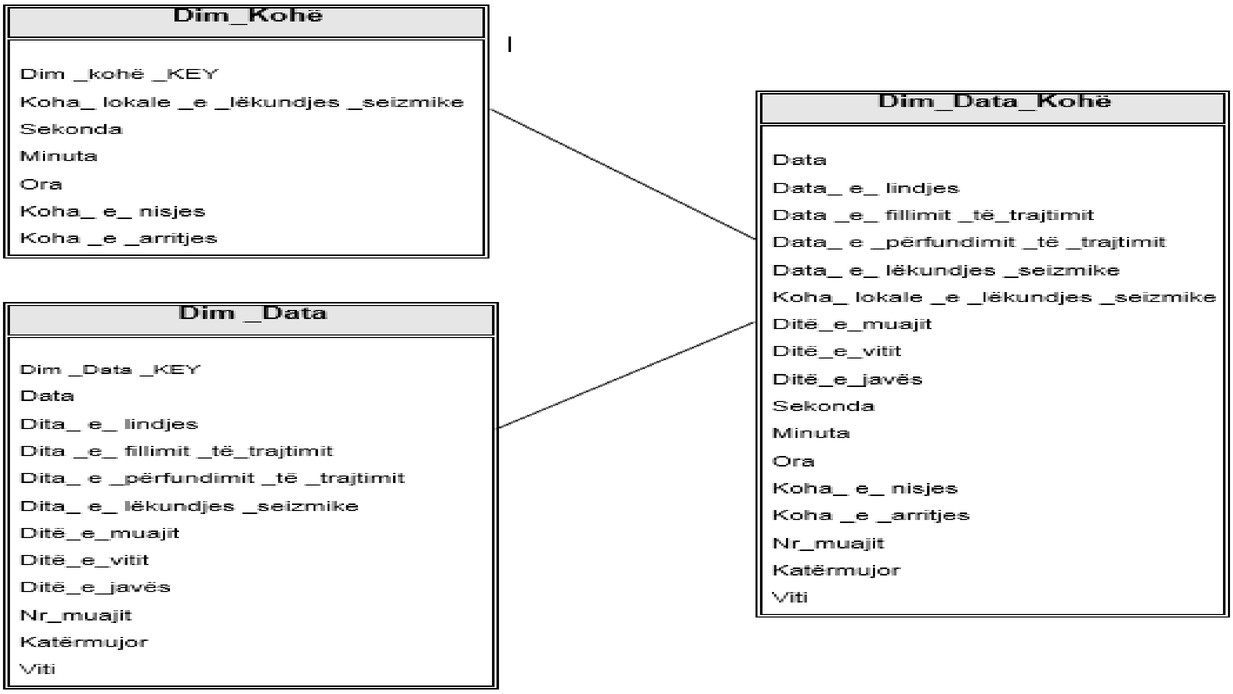
## DIMENSION TABLES

Dimension tables are integral companions to a fact table. The dimension tables contain the textual descriptors of the business. In a well-designed dimensional model, dimension tables have many columns or attributes. Dimension attributes serve as the primary source of query constraints, groupings, and report labels. In a query or report request, attributes are identified as the by words. Dimension table attributes play a vital role in the data warehouse. Since they are the source of virtually all interesting constraints and report labels, they are key to making the data warehouse usable and understandable. In many ways, the data warehouse is only as good as the dimension attributes. The best attributes are textual and discrete. Attributes should consist of real words rather than cryptic abbreviations. Typical attributes for a product dimension would include a short description (10 to 15 characters), a long description (30 to 50 characters), a brand name, a category name, packaging type, size, and numerous other product characteristics. Although the size is probably numeric, it is still a dimension attribute because it behaves more like a textual description than like a numeric measurement. Size is a discrete and constant descriptor of a specific product. Sometimes when we are designing a database it is unclear whether a numeric data field extracted from a production data source is a fact or dimension attribute. We often can make the decision by asking whether the field is a measurement that takes on lots of values and participates in calculations (making it a fact) or is a discretely valued description that is more or less constant and participates in constraints (making it a dimensional attribute). For example, the standard cost for a product seems like a constant attribute of the product but may be changed so often that eventually we decide that it is more like a measured fact. (Kimball 2008).Earlier, there was a tradition to follow all three phases of modeling while building a database. The first model that should be built is the conceptual one, which is focused towards business to collect its requirements in a high level. It is generally represented by a scheme and should be independent from the type of the database used. The next model is the logical one, which must describe the exact scheme used by the database. And lastly there is the physical model, which is composed of a compound of scripts which generate and create tables, indexes, links, etc. In nowadays, many technicians merge all three models in a single phase. They simply use one model (ER) and convert it directly in a compound of database objects. However, a DWH is different from other types of databases and as a result, it is recommended to use the trilogy conceptual/logical/physical. (Juliana 2012)

## CONCEPTUAL MODELING

During the conceptual modeling, a special place was taken by the time element. The time dimension is part of any possible analysis, it may be necessary up the level of seconds for detailed analysis (Juliana 2014). Even though the date dimension is the most important dimension of time, we also need a monthly dimension where the time span of the table of facts is one month. In other environments, we might need to build weekly, quarterly or yearly dimensions and if there are tables of facts in each of these spans (Kimball 2004). Based on the multi-dimensional models, the time element was added as a dimension that changes quite often and must be partitioned. According the above understanding, the dimension date-time should be partitioned in two tables of our scheme (Juliana 2014).

Fig: Partitioning of dimensions in the scheme



In this case, it is useful to make a dimension with components of minutes or seconds of the exact time, because the calculation of time intervals in the table of facts records becomes very complicated when we try to deal with special dimensions of day and time-of-day (Kimball 2004).

This prepared document will represent the core of this DataMart and will serve as a vocabulary (metadata). After finishing with the conceptual model we will continue with the building of the logical model (Juliana 2014).

## LOGICAL DATA MODEL

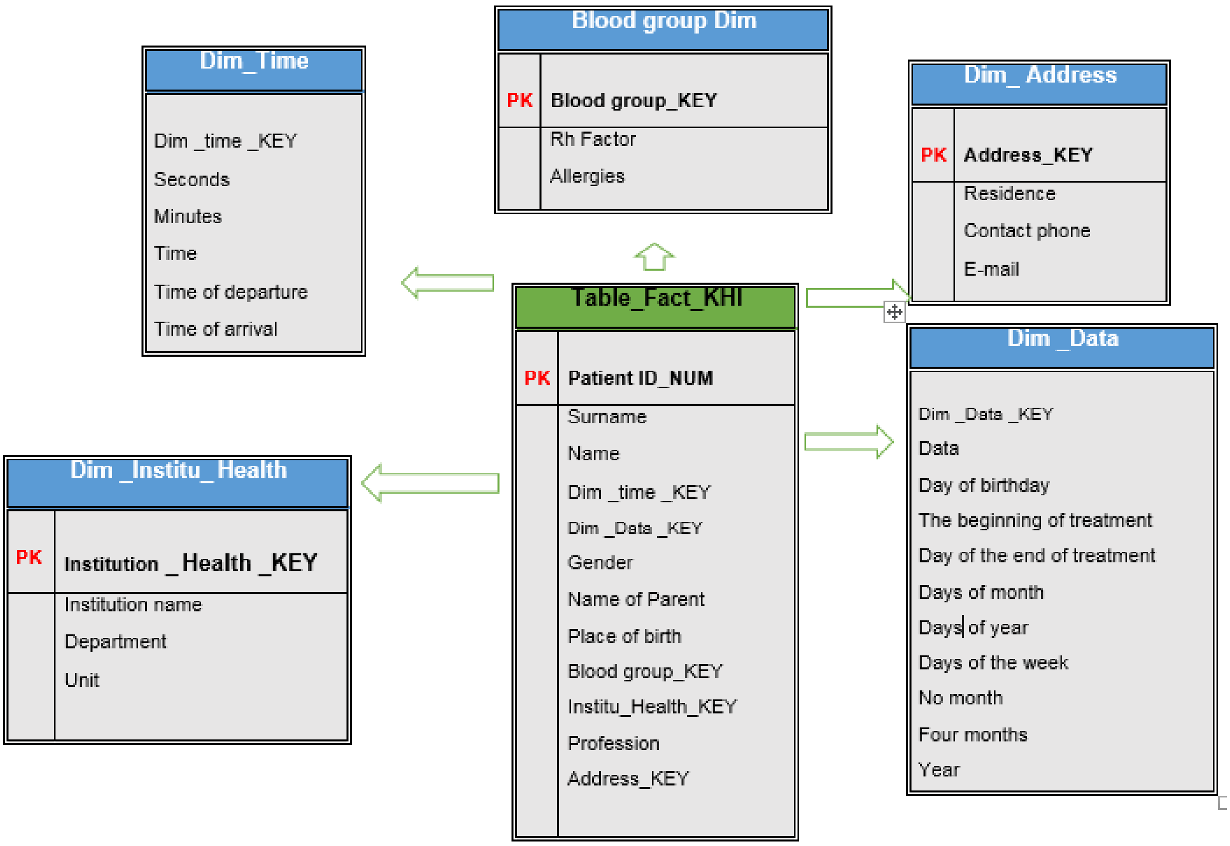
Logical modelling transforms the conceptual data model into a dimensional model, commonly known as a star schema. It consists of a large table of facts (known as a fact table), with a number of other tables surrounding it that contain descriptive data, called dimensions .Logical data model includes the following features:

- Includes all entities and relationships among them.
- All attributes for each entity are specified.
- The primary key for each entity specified.
- Foreign keys (keys identifying the relationship between different entities) are specified.

At this level, the data modeler attempts to describe the data in as much detail as possible, without regard to how they will be physically implemented in the database. Each of dimension names shown on the dot diagram usually becomes a separate dimension table in a star schema. (Camilovic 2009)

Except the dimension structuring, there also must be built fact tables. Generally this type of table contains all ‘foreign keys‘, and all the numerical measurements called facts. In the case of the facts table of modeling there will be only one numerical measurement which has to do with the call duration (Juliana 2012).
The DataMart scheme of IMK is described below:

Picture 3. The DataMart scheme of KHI



## Physical Data Model

At this level, the data modeler specifies how the logical data model will be realized in the database schema. The objectives of the physical design process do not center on the structure. The modeler is more concerned about how the model is going to work than on how it is going to look. According to Ponniah, there are several steps in the physical design process for a data warehouse:

• Developing standards. The standards range from how to name the fields in the database to how to conduct interviews with end users for requirements definition. The standard for naming the objects takes on special significance, because the usage of the object names is not confined to the IT specialists. The users also refer to the objects by names, when they create and run their own queries. This is why the name itself must be able to convey the meaning and description of the object. It is also necessary to adopt effective standards for naming the data structures in the staging area, as well as all types of files (not only data and index files, but also files holding source codes and scripts, database files and application documents).

• Creating aggregates plan. In this step, the possibilities for building aggregate tables should be reviewed. It is possible that many of the aggregates will be presented in the OLAP system. However, if OLAP instances are not for universal use by all users, then the necessary aggregates must be presented in the data warehouse. There are two primary factors that should be considered when making a decision about the aggregates plan. First, the business users' access patterns have to be taken into account – data that they are frequently summarizing on the fly should probably be presented in the data warehouse. Second, the statistical distribution of the data should be assessed (e.g. how many unique instances exist at each level of the hierarchy, and what's the compression in case of moving from one level to the next)

• Determine the data partitioning schema. Partitioning options for fact tables, as well as dimension tables, should be considered. The data warehouse usually holds some very large database tables. The fact tables run into millions of rows, and many dimension tables (e.g. customer tables) contain a huge number of rows, as well. Having tables of such extremely large sizes faces certain problems. Loading of large tables and building indexes takes excessive time. Queries also run longer, backing up and recovery of huge tables, too. This is why the partitioning is important. Partitioning means deliberate splitting of a table and its index data into manageable parts. (Kimball 2004)

## CONCLUSIONS

Taking into consideration the logic of building the DataMart of Emergencies Management Agency (EMA) and the generated data it was noticed that the flexibility that offers such a model in generating analysis is high. This model overcomes the difficulties of the data volume. The application of Data Warehouse methods was one of the most critical tasks for the integration of many sources in one centralized system (as were the data of *Kosovo Health Institution –KHI*, Kosovo *Seismologic Institution - KSI and Kosovo Hydro-Meteorological Institution - KHI*). During this phase, we encountered more difficulties in the standardizing and structuring of data.

In nowadays, the data integration from many sources allows for the gathering and analysis of massive data and detailed data records. This is the hottest topic in exploring knowledge in civil emergencies management. For this reason, the DataMart modeling for this data is a very interesting topic. This DataMart can be extended to include other additional information which would further help the analysis and management decision making.

## REFERENCES

Arwa Abdullah J (2011): Transitioning a Clinical Unit to a DataWarehouse.
Camilovic D (2009): A Call Detail Records Data Mart: Data Modelling and OLAP Analysis
Jeffrey A. H, Ramesh.V, Heikki .T (2011): Modern Database Management
Juliana L (2014): The study of data warehouse systems and the construction of a reporting model based on business intelligence technology, Tirana, Albania
Juliana L, KIKA .A (2012) : Modeling a DataMart-of CDR
Kimball & Caserta (2004): The Data Warehouse ETL Toolkit ,Wiley
Kimball.R (2008): The Data Warehouse Lifecycle Toolkit, Second Edition
Shaker. A *(2011).*A proposed model for data warehouse ETL processes
Google.2010."Strategy for health information system in Kosovo 2010 – 2020"

http://www.kryeministriks.net/repository/docs/STRATEGJIA__PER__SISTEMIN_E_INFORMIMIT_SHENDETESOR_NE_KOSOVE_2010_-_2020.pdf (Click:15.09.2015)

Google.2011.”Kosovo Flood Risk Management Framework”http://www.kryeministri-ks.net/tfu/repository/docs/110421_Kosovo_Flood_Management_Framework_Alb.pdf Click:15.09.2015)