Business Intelligence and Visual Analytics-Based Strategy for disaster risk observatory development

Estrategia basada en inteligencia empresarial y análisis visual para el desarrollo del observatorio de riesgo de desastres

E. Avendaño[1], A. Moreno[2], G. Sanchez[3], M. Villamil[4]

[1] *M.Sc., Facultad de Ingeniería, Universidad del Magdalena, Grupo de Investigación y desarrollo en sistemas y computación, Santa Marta, Colombia. eavendano@unimagdalena.edu.co*

[2] *PH.D., Departamento de Ingeniería de Sistemas y Computación, Universidad de los Andes, COMIT-Comunicaciones y Tecnología de la Información, Santa Bogotá, Colombia. dar-more@uniandes.edu.co*

[3] *PH.D., Facultad de Ingeniería, Universidad del Magdalena, Grupo de Investigación y desarrollo en sistemas y computación, Santa Marta, Colombia, gsanchez@unimagdalena.edu.co*

[4] *PH.D., Departamento de Ingeniería de Sistemas y Computación, Universidad de los Andes, COMIT-Comunicaciones y Tecnología de la Información, Bogotá, Colombia. mavillam@uniandes.edu.co*

Cite this article as: E. Avendaño, A. Moreno, G. Sanchez-Torres, M. Villamil. "Business intelligence and visual analytics-based strategy for disaster risk observatory development", Prospectiva, Vol 18, N° 1, 13-23, 2020.

Recibido: 10/10/2019 / Aceptado: 19/11/2019

https://doi.org/10.15665/rp.v18i1.2258

ABSTRACT

In this article, we present a development strategy for disaster risk observatories, integrating business intelligence and visual analytics. Analysis and contextualization at a global level are carried out from the perspective of policies, tools, and scientific literature in the context of disaster risk reduction. The strategy integrates the Kimball methodology for the construction of business intelligence (BI) projects and the visualization framework of Tamara Munzner to strengthen the decision-making process and make it flexible and faster, satisfying the needs of business users. Considering disaster risk management as a process allows for the identification of the starting point and the processes that follow, resulting in a quick initiation of the project lifecycle. In addition, including the visualization framework, along with the BI methodology, facilitates the development of analysis tools for solving end-user issues.

Key words: Business Intelligence, Disaster Risk Management, Disasters Risk Observatory, Visual Analytics.

RESUMEN

En este artículo, se presenta una estrategia para el desarrollo de observatorios de riesgo de desastres, integrando inteligencia empresarial y análisis visual. El análisis y la contextualización a nivel global se llevan a cabo desde la perspectiva de políticas, herramientas y literatura científica en el contexto de la reducción del riesgo de desastres. La estrategia integra la metodología Kimball para la construcción de proyectos de inteligencia empresarial (BI) y el marco de visualización de Tamara Munzner para fortalecer el proceso de toma de decisiones y hacerlo flexible y más rápido, satisfaciendo las necesidades de los usuarios empresariales. Considerar la gestión del riesgo de desastres como un proceso permite la identificación del punto de partida y los procesos que siguen, lo que resulta en un inicio rápido del ciclo de vida del proyecto. Además, incluir el marco de visualización, junto con la metodología de BI, facilita el desarrollo de herramientas de análisis para resolver los problemas del usuario final.

Palabras clave: Inteligencia de Negocios, administración de riego de desastres, observatorio, analítica visual.

1. INTRODUCTION

Population growth occurs exponentially, and the concentration of inhabitants in urban areas is increasing significantly [1], [2]. It is predicted that 66% of the world population will live in cities by 2050 [2]. Similarly, the impact, magnitude, and frequency of disasters have also increased. On average, 240 million people were affected by natural disaster events worldwide from 2000 to 2005, with estimates of more than 80,000 fatalities and costs of almost US $80 billion [3]. It is possible to avoid or reduce this impact by establishing adequate policies and programs, defining mitigation procedures, and establishing response mechanisms that are effectively integrated into urban development plans [3]. Thus arises the challenge of building policies for sustainable growth that consider natural disasters in order to avoid human losses and material costs.

In Colombia, the number of emergencies linked to natural disasters is increasing. The social crisis is further intensified due to the loss of lives, properties, and municipal infrastructures, among other things [4], [5]. The establishment of prevention policies should lead to an increase in disaster risk studies. In [6], it is stated that natural disaster risk studies must be carried out as an analysis in which the risk depends on multiple social and economic variables. The threat of natural disasters in the territory, as well as the social vulnerability of the area, should be taken into consideration when calculating disaster risk. Therefore, risk reduction should be explicitly introduced as an objective of social development addressed to the improvement of life quality and social welfare [7].

In decision-making and planning, disaster risk management (DRM) can reduce the impact caused by natural disasters. The starting point for disaster risk reduction is the knowledge of the threats and vulnerabilities, which leads to the formulation of decisions based on this knowledge [8], [9].

The study of the associations of multiple variables and various sources of information is the foundation of disaster risk analysis: regulations, maps, historical data on disasters, and socioeconomic conditions, among others. Consequently, the problem has multiple dimensions that magnify its complexity in real environments.

In order to prevent disasters, it is necessary to understand the relationship between risk, threat, and vulnerability, which facilitates risk assessments [10].

Observatories are platforms that allow the selection, storage, integration, and presentation of different data sources, providing information to users in a specific knowledge domain [11]. Therefore, they constitute an essential tool for disaster risk assessment due to their organization of related historical information. The creation of risk observatories will facilitate decision-making for risk knowledge, management, and reduction.

The United Nations Office for Disaster Risk Reduction (UNISDR) [12], defines Disaster Risk Management (DRM) as the administrative direction process and the operating capacities for establishing policies for reducing threat impact and disaster likelihood. In order to prevent and reduce the effect of disasters, it is necessary to understand the interactions between risks, threats, and vulnerabilities [9], [10].

We highlight two key concepts from the formal definition: "threat" and "vulnerability." A threat can be understood as the potential danger of a physical event happening and causing human casualties, material losses, or environmental damages. As for vulnerability, law 1523 [13] of the Republic of Colombia defines it as:

"Physical, economic, social, environmental, or institutional susceptibility or fragility that a community has of being affected or of suffering adverse effects in the case of a hazardous physical event. It corresponds to the predisposition of human beings or their livelihood to suffer losses or damages. It also comprises the possibility of their physical, social, economic, and support systems being affected by hazardous physical events."

Next, when we combine the danger of an event happening with the fragility of a community, we find ourselves facing a disaster risk [14].

DRM can be divided into three sub-processes: risk knowledge, risk reduction, and disaster management. It must be noted that they are not isolated. Instead, there is a dependency between them, and they are continuously executed. This means that in order to comprehend risk management, risk reduction and knowledge must have been addressed previously [13], [15].

A) Tools for disaster risk management

Tools in the literature address different disaster risk management processes: emergency and disaster knowledge, reduction, and management.

Inform is the first unbiased, transparent, and global tool for understanding disaster risk. It is useful for evaluating disaster risk, and it is based on a three-dimensional methodology: threat and exposure, vulnerability, and lack of response capability [16].

Other tools employed internationally are the *CAPRA* (Central American Probabilistic *Risk* Assessment) [17] platform and the tools developed by the Centre for Disaster Management and Public Safety (CDMPS) [18]. The former combi-

nes threat, exposure, and physical vulnerability information for calculating risk either jointly or sequentially. The latter is a research center located in Australia and dedicated to multidisciplinary risk management research. CDMPS has several projects which have spawned a series of disaster management applications, including *The Australia Disaster Management Platform (ADPM)* [19], *An Intelligent Disaster Decision Support System (IDDSS)* [20], *Risk Finder* [21].

In Colombia, the SIRE [22] (Sistema de Informacion para la Gestion del Riesgo y Cambio Climatico) can be highlighted. This website is a platform where applications focused on risk management, and climate change can be accessed. Within SIRE, there is the Risk and Climate Change Management District System (SDGR-CC) [23], which is related to the following processes: Risk knowledge, Risk reduction, Emergency and disaster management, and Climate change relief and adjustment. The SIRE website and that of the District Institute for Risk Management and Climate Change (IDIGER) are intertwined, and this is remarkable because the latter website [24] provides additional risk scenarios such as floods, slope movements, seismic movements, significant public events, and forest fires.

Another tool employed in Colombia is the Geoportal of the Colombian Geological Service [25]. This tool displays information related to mineral resources, basic geosciences, geothreats, laboratories, and information management.

B) Works related to Disaster Risk Management

Perez, in [26], argues that part of the population in the Argentinian city of Neuquén inhabits high-risk areas. Besides, he notes that the problem is complex. He also suggests cartography as an appropriate tool for assessing risk in each sector of the city. In order to identify high-risk zones, Perez employs the Social Risk Theory [26], which takes into account four basic dimensions: dangerousness, social vulnerability, exposure, and uncertainty. He also uses maps to display the first three dimensions of the theory. In [27], Colombian situation is described and links its geography to the various risks the country is exposed to. He notes that population growth, human settling in risk zones, and fast urbanization without proper planning increase population vulnerability and man-provoked threats. He also suggests that the biggest issue is not the fact these phenomena occur, but instead that there are no clearly-established mechanisms for counteracting them.

We propose a development strategy for a disaster risk management observatory, which will provide access to historical data of threats and affected populations. This observatory will serve different actors, such as first-response agencies, decision-makers, and the general community, allowing them to view their territorial conditions and the possible risk scenarios using different display options. The information available to them will be based on historical data. Additionally, it is useful for decision-makers to have a guide to creating intervention programs and disaster risk reduction.

2: METHODOLOGY

Strategy for developing disaster risk observatories

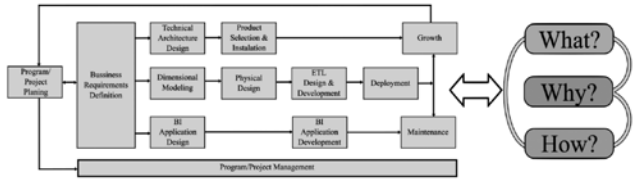
For this proposal, we adapted the life-cycle methodology of Kimball and integrated it with Munzner's Visual Analytics Framework [28], [29].

Figure 1 depicts the methodology of Kimball (left side) and the kernel of Munzner's framework (right side). It is relevant to propose the integration of the two proposals, as they share a common goal: supporting decision-making. The need for employing a visualization framework arises when there are multiple ways to represent data. Previous work on how to represent a given problem should be studied, and researchers should review the documentation of prior proposals in order to understand them. The three research questions proposed by Munzner to streamline this process are [29]:

- *Why?*. This refers to the issues that must be addressed with the visualization tool.
- *What?*. This refers to the data that needs to be visualized.

How?. This refers to the encoding of the visualization (idioms and interactions).

Figure 1. Kimball's methodology and Munzner's visualization framework [28], [29].



The integration strategy that will facilitate the development of disaster risk observatories is depicted in Figure 2. Each box represents a task, and the arrows indicate the next task to be carried out.

The first task we need to carry out is *Observatory planning*, whose goal is to understand disaster risk management and the processes that govern it in order to define the scope of the observatory. For this purpose, we need to understand what the DRM is and what its legal framework is. This allows us to define the variables and to limit the scope of the observatory to a specific domain.

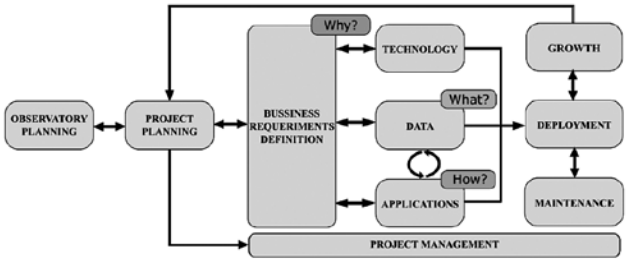
It is worth noting that, unlike in Kimball's proposal, process prioritization is not necessary for our proposal. The reason is that DRM processes are sequential: they start with risk

knowledge, followed by risk reduction, and finally disaster response.

The second task is *Project planning*. Here, a schedule with the activities is created, roles are defined and associated with each activity, and the scope of the project is established. This is done by using a spreadsheet program or any tool for generating schedules.

The next task is *Project management*, which is parallel to project planning. Project management allows the state of the project to be monitored continuously and enables communication between end-users and developers. Defining business requirements is crucial for the execution of the project. This is because it is a prerequisite for carrying out a set of parallel design and development tasks (technology, data, and application). For these tasks, new interviews are conducted with a higher level of detail.

Figure 2. Proposed strategy for disaster risk management observatories. It integrates Kimball's methodology with Munzner's framework.



The Framework uses *Why* to identify the tasks in abstract form by separating them into *actions* and *targets*. An example of a *task*, in these terms, would be to discover the distribution of neighborhoods with a given risk level. The verb *discover* refers to the *actions* and *distribution of neighborhoods* refers to the *targets*. Three tasks can be started in parallel after *defining the business requirements*. The first one is the data task. This task is comprised of several activities: profiling, exploration, dimensional modeling, physical design, and ETL (extract, transform, and load) design and development.

The profiling and exploration activities consist of reviewing the data resources to understand them and to identify possible quality findings, such as incomplete or null data, out of range data, among others.

The creation of a dimensional data model requires a paradigm shift, as it is different from 3NF (third normal form) relational database design. Regardless, Kimball's group provides a sizable amount of documentation, which contains solutions to several modelling problems [30].

In order to design the dimensional model, a high-level model based on Kimball's bus matrix is first created. Then the level of detail that will be used to store the records, also known as granularity, is defined. Lastly, dimensions, measures, and facts will be selected [31]. As for the physical design, the main goal is to select the attributes that will belong to each table and to choose how dimensions will be versioned. This means that some dimensions may change in time. There exist several options to address this issue [30]:

- One, it could be that we are not interested in that change of the name product, and we do not change anything in our dimensional model. That change will not affect us, and we will continue to show the first name with which the product record was created. This is known as a type 0: *slowly changing dimension* (SCD).
- Two, we are interested in doing the *update,* and the new name will appear both in the news reports and in the old ones. This is known as a type 1: *slowly changing dimension*.
- Three, we create a new product record with the new name, but we use date fields to indicate from which and until which date registration is valid. This form of change is known as a type 2: *slowly changing dimension.*
- Four, in this option, we are interested only in version tracking, whereby a new column is created in the product dimension called "current product name". This form of change is also known as a type 3: *slowly changing dimension*.
- Five, in this option, we can create a new table where all the historical changes that a product has had are recorded. This form of change is known as a type 4 *slowly changing dimension.*
- Six, this is an option that combines type 1, 2, and 3 SCDs. It consists of adding a column with the current field and handling register validity dates. This solution, although much more complex, allows all historical changes to be kept and the current version to be known. This form of change is known as a type 6 SCD because is the sum of the types 1, 2 and 3.

As for ETL design and development, it must be noted that it is divided into three sub-tasks: extract, transform, and load. These sub-tasks are usually under-estimated. However, there is also a lot of documentation and many tools for carrying them out [32]. If the dimensional model is the heart of the project, the ETLs are its veins. Correctly developing ETL is crucial, since if they do not work, there will not be any data for the model.

It is necessary to analyze the double arrow that exists between Data and Requirement definition (see Fig. 2). This arrow means that if we find a data quality finding which

affects the analytics requirement when carrying out data exploration or profiling, then we must return to the requirement definition task and redefine the tasks. We can also use Munzner's framework for data activities as it allows us to identify data types and their semantics.

Understanding the kind of data that is handled allows us to determine which kinds of analysis and processing will be necessary for visualizing them [29]. With *what* abstraction, we can understand the data and dataset types.

Another task that must be carried out is the Technology task. This task is centered on the architecture design activities for the observatory and the selection of the tools to be employed. At this point, the architecture for the observatory is defined (see Fig. 3). There are four essentials elements in the observatory architecture: Data sources, sources reliable of data in many formats; Repository, is the place where the data is modeled and stored; information representation, is comprised of the apps used for showing the information needed for the end-users. For the tool *selection* activity, business restrictions must be taken into account, as this will allow us to determine whether to use a proprietary software solution such as Tableau or to employ open-source solutions such as *Metabase* or *Pentaho Business Analytics*. The Application task involves designing and developing the applications that allow us to represent data and to generate reports. These make up the interface all the users interact with. This task is supported by Munzner's framework through the *How?* question, which refers to how the visualization will be carried out, the legends, how they will be organized, and the interactions with the end-user.

There are two arrows between the Application and Data tasks. These indicate there is continuous interaction among them. This happens because there is a high demand for methodological and framework flexibility, where end-user needs are a priority.

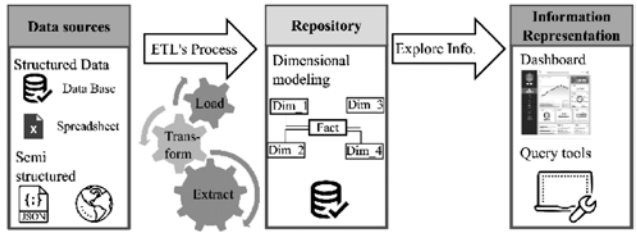
The signaled interrelation happens through the interaction with the end-users and their analysis tasks. Thus, the storage model and the user interface are evaluated continuously. One of the challenges of this strategy is to define when it is necessary to freeze a model or an interface as not to spend too much time carrying out changes that drain out resources and slow down project development.

If the project monitoring done through the Project management task were successful, Deployment would be simple as every piece must have functioned properly. The Maintenance tasks reference the necessary adjustments that come up after implementation. As for Growth, it happens when users request changes after using the application or when the next DRM process is carried out.

3. CASE STUDY

In order to verify the functionality and veracity of the proposed strategy for risk management observatories, we carried out a case study in the city of Santa Marta. We present next the main results from implementing the strategy.

Figure 2. Proposed strategy for disaster risk management observatories. It integrates Kimball's methodology with Munzner's framework.



A) Observatory planning

In order to address the project requirements and to understand the essence of a risk observatory, it is necessary to investigate the existing definitions for risk and the methodologies for its evaluation, as well as the existing legal framework. This will be the basis for understanding the data that should be obtained and the results to be presented.

As a result of this research, we know that the Colombian normativity often depicts a generic definition of risk: a combination of threats and vulnerabilities. In addition, we employ the semiquantitative methodology proposed by the PNUD and the UNGRD [15] for evaluating risk. We took this decision based on the research we carried out as well as the first interview with the business user: the director of the office for disaster risk management and climate change.

In the aforementioned methodology, the following equation for threat analysis employed [15]:

$$Threat(A)=Intensity(I)+Frequency(F)+Affectedterritory(T) \quad (1)$$

Intensity is the quantitative and qualitative measure of how severe a phenomenon is in a specific territory. It is rated as high, medium, and low depending on the number of human casualties, the effect on natural resources, the suspension of public services, and the number of destroyed/damaged housing units. Frequency is how often threatening phenomena occur in the territory. It is rated as high, medium, and low, depending on how many times the event happens in a period defined in years.

The territory is the area comprised of portions of land, rivers, seas, gulfs, ports, canals, bays, among others, and which can be affected by threatening phenomena. Affected territory is rated as high if more than 80% of the territory is affected, medium if the amount is between 50 and 80%, and low if the

affected portion is of less than 50%. As for *Vulnerability,* it is considered as a factor for risk analysis and is comprised of 4 kinds of subfactors: *Physical, Economical, Environmental and Social.*

Physical vulnerability is linked to the type and quality of the materials employed for constructing housing units, commercial/industrial establishments, public service buildings (hospitals, schools, governmental institutions), and socioeconomical infrastructure (hydroelectrical plants, roads, bridges). Another important aspect is the characterization of the territory itself and to determine whether populational centers are close to geological fault lines, rivers, and coastlines since such conditions can significantly increase vulnerability. The variables to be analyzed include building age, construction materials and conservation state, fulfillment of current normativity, geological features and kind of soil, and the location of buildings relative to water source exclusion zones and other risk zones.

Economic vulnerability is determined by the income level and the capacity of the population to fulfill basic needs. Variables to be analyzed include poverty level and food safety, income, public services accessibility, and labor market accessibility. Environmental vulnerability is the degree of endurance of the environment and the living beings belonging to it when exposed to climatic variability. The variables to be analyzed are atmospheric conditions, air composition and quality, water composition and quality, and environmental resource conditions. Social vulnerability is analyzed based on how organized and participative a community is, and how well it can prevent and respond to emergencies. The variables to be analyzed are organization, participation, collaboration between community organizations and institutions, and community risk knowledge. We must note that each of the vulnerability variables is in the 1 to 3 scale. In order to compute total vulnerability, we employ the following equation:

$$Vt = Vf + Va + Ve + Vs \quad (2)$$

By assigning values for every variable in the equation, we obtain an aggregate value for scoring vulnerability as low, mid, or high as Low for values in range 16-26, Mid for values in range 27-37, and High for values in range 38-48.

As risk analysis is based on the identification and evaluation of probable damages and losses as a consequence of the impact of a threat on a vulnerable social unit [33]. There is a descriptive model for scoring risk based on a 2D matrix of threats and vulnerabilities [15] (see Table 1).

After quantifying total threat and vulnerability scores, we place the resulting threat score on the y-axis and the resulting vulnerability score on the x-axis. We then look for the intersection of the values to define the risk level.

Having the scoring methodology for the risk we establish the first restrictions. We can then understand that disaster risk management can be handled as a process with three subprocesses with a pre-established order. According to Colombian law 1523, "risk knowledge is the process comprised of scenario identification, risk analysis and evaluation, risk monitoring and tracking, as well as the associated factors which can serve as an input for risk reduction and disaster management" [13].

Table 1. Risk score

	Low Vulnerability	Mid Vulnerability	High Vulnerability
Low Threat	Low Risk	Low Risk	Mid Risk
Medium Threat	Low Risk	Medium Risk	High Risk
High Threat	Medium Risk	High Risk	High Risk

By understanding this law, we can start planning observatory development. The first step for risk management is knowledge, so the first process to be carried out is the "Knowledge" process, which is related to risk evaluation [15].

After clearly understanding the relevant process, we carried out interviews with risk management entity administrators in the city of Santa Marta, which allows us to identify the most relevant analytical issues in the risk knowledge process. This allows us to build the Kimball bus matrix (see Table 2) for the disaster risk knowledge process

The first column of the matrix includes the analysis issues related to the risk knowledge process. The first row includes the dimensions, which are the criteria that can be employed to the group, summarize, and query the data to be analyzed. In addition, the dimensions can be used as metrics for assessing the analysis issues.

The analysis issues belonging to the risk knowledge process are:

Neighborhoods exposed to the risk. This allows us to establish the geographical area to be analyzed.

Disabled persons exposed to the risk. These are the persons that are the most vulnerable, due to conditions such as reduced or null mobility and their capacity to see or to hear.

Conditions of the housing units exposed to the risk. This issue refers to aspects such as the physical and social configuration of housing units. More specifically, it assesses a number of family members, ages, income, schooling, number of rooms, access to utilities, among others.

After identifying the process, defining the methodology, and establishing the main analysis issues, we finish the observatory planning task and start the project planning and management task.

Table 2. Kimball's bus matrix

DRM process – analysis issues.		DIMENSIONS											
		Person	Time	Threat	Vulnerability	Risk level	Location	Housing unit	Age	Income	Disability	Schooling	Climate
Risk knowledge	Neighborhood exposed to the risk.		■	■	■	■	■	■					■
	Disabled persons exposed to the risk.	■	■	■		■			■	■	■	■	
	Conditions of the housing units exposed to the risk.		■	■		■	■	■					

B) Project planning and Project management

In this step, we name the project, and it will form the first iteration of the "Analysis and evaluation of multidimensional risk" cycle, the schedule of the following tasks, and its roles are defined (see Table 3). With table 5 we identify the roles associated with each task, so the director can carry out monitoring more efficiently. In this task, several project restrictions are defined, which were based on meetings with the risk manager of the city of Santa Marta:

- No proprietary software can be employed.
- Only data from the urban area of Santa Marta can be employed.

The next task is to define business requirements.

C) Business requirements definition

In these tasks, meetings and interviews should be carried out with end-users in order to better understand the decisions they make and the steps they take for that purpose (data sources, treatment). As for the visualization framework, the goal is to define human tasks, in terms of business intelligence, the goal is to define analytics requirements.

The principal requirements we identified were:

- Risk analysis and evaluation in a neighborhood for a specific period.
- Identification of meteorological conditions and the neighborhoods affected by such conditions.
- Analysis of the physical configuration of homes in a neighborhood for a specific period.
- Analysis of the vulnerabilities in a neighborhood for a given neighborhood.
- Analysis of the distribution of risk, threat, and vulnerability levels in a per-neighborhood basis for a given period.
- Identification of persons with reduced mobility in a neighborhood.
- Discovery of the income distribution of people for a given period.

Table 3. Activities associated with each role

Role / Task	Project planning	Requirements definition	Data	Applications	Technology	Deployment	Maintenance	Growth
User			■	■			■	■
Observatory director	■	■						■
Project Director / Solution architect	■	■	■	■	■	■	■	■
Business analyst		■	■	■				■
Data architect			■		■	■		
ETL architect			■		■	■		
BD developer				■				
Web developer				■				
SQA				■			■	
Security administrator						■		

D) Data

In this task we review, profile, and explore data sources. In addition, we design the database and define the ETL techniques to be employed.

1. Data comprehension

There are several tools that can be employed for carrying out this task. Specifically, we abstract data using *what* component of Munzner's framework. As for the data itself, the sources are diverse, as well as the formats. The sources were DADMA (Departamento Administrativo Distrital del Medio Ambiente), Santa Marta Planning Secretaryship, and Climate API.

DADMA supplied a .kmz file that contained the geometry of the neighborhoods of Santa Marta. This kind of file is usually employed for applications such as Google Earth.

The main feature of the SISBEN card is that it allows to collect of housing units, home, socioeconomic, health, education, and work data from people. From the housing unit data, it is possible to know physical properties such as the kind of soil, the kind of construction materials employed, the number of rooms, the conditions of toilets, among others. As for home data, it is possible to identify home appliance possession and food preparation conditions. As for persons themselves, it is possible to know whether they are disabled, their age, their sex, their schooling, their income, among others. SISBEN data includes both urban and rural homes. The data we employed includes the results of surveys carried out from 2015/01/01 to 2016/12/01.

The climate API data was retrieved in an unstructured. json format. It included historical data from 2015/01/01 to

2016/12/31 for the city of Santa Marta. It also included forecast data from april 2017 to june 2017. The variables we empluyed are Rain probability, Precipitation in mm, Temperature, Wind speed and direction, Atmospheric pressure, and Visibility.

2. Data profiling

The goal of data profiling is to verify data quality. For this purpose, we employed two tools on the SISBEN data:

- Pandas_profiling Python library: With this library we can present a preliminary summary of the data, classifying variables as numerical or categorical. Furthermore, we were able to identify variables that were constant or that presented very little variability. The SISBEN records are comprised of 121 variables, but not all of them are necessary for the project, so in the modeling task it is necessary to select which of these are to be employed.
- DQAnalyzer: This tool allows profiling to be carried out in a different manner. With DQAnalyzer, masks of string variables can be analyzed. This was particularly helpful for handling address data. Mask analysis shows in which position of a string there is a number, a letter, or a special character. Based on this, we discovered only 24% of the records had a valid address format, causing us to discard this variable for further analysis.

3. Data exploration

Exploration allows us to better understand data, as it allows us to experiment with it to discover relevant information.

4. Dimensional model design

The dimensional design is based on the bus matrix built in the planning task. This matrix allows us to create a high-level diagram that depicts the dimensions and facts. From this activity, we obtained three fact tables: climate, homes, and risk. Climate is described by time and by wind direction. We discarded several fields associated with wind direction records, such as name, abbreviature, and code, as they were not required for our purposes.

Fig. 4 depicts the risk fact and its dimensions: threat, vulnerability, location, and time. The risk fact comprises the detail of every variable and vulnerability variable to have occurred in a specific place, which in this case is each neighborhood in the city. In addition, we designed the home fact with its dimensions: housing unit, time, and person. This model allows us to analyze homes from the point of view of the persons that inhabit them and the physical features of the housing unit, such as access to utilities, the number of rooms, etc.

5. Physical-dimensional design

In the physical design, it is necessary clearly establish which attributes are going to be included in each table. As part of this activity, the attributes are selected and the kind of versioning to be employed for each dimension is defined. For our dimensional model we decided to implement slowly changing type 2 and type 6 dimensions. SCD type 6 was employed for the housing unit and person tables. Next, we elaborate on how the physical model was implemented for the housing unit dimension.

If a person updates their information after moving to a different housing unit with a different address in a new neighborhood, what needs to be done is to aggregate the previous and current fields for both the address and the neighborhood so historical data for both fields can be preserved.

6. ETL design and implementation

We designed and implemented 3 ETL processes, one for climate data, other for risk data, and another for SISBEN data. Next, we explain the ETL process for SISBEN data. This process was, too, the most complex and interesting, as was implemented using type 6 SCD.

The general steps:

1. We first clean the staging tables, which are intermediate tables that can be used to carry out transformations.
2. Since the SISBEN data is loaded from files, we verify that the file exists in the specified route. If it does not exist, the exception is logged and the job execution is canceled. Otherwise, the name of the file is retrieval.
3. We carry out a new transformation where we obtain the filename and we filter data to only insert data from the Santa Marta urban zone into the staging table.

3.1. In this transformation, we also store a record of the file which is going to be loaded into the data mart [28].

4. Two stored procedures are executed in order to transfer data from the staging table and the housing unit and person dimensions.

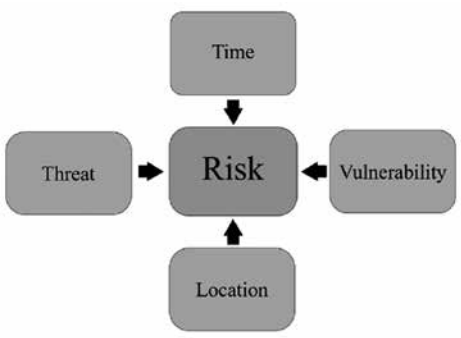
4.1. In order to carry out this step, we employed an algorithm based on finding the identifiers of the modified records. Then, we update the records. Then, we insert new records.

5. Finally, we fill the facts table.

Once this is done, we update the file load control table and end the execution.

After finishing the data task, the next task to be carried out is the technology task.

Figure 4. High level logical design for the risk fact and its dimensions.



E) Technology

In this activity, we considered the business restriction which states no proprietary software can be employed. We determined the following tools were appropriate for our purposes:

- For developing the risk data management web application, we employed Ruby on Rails (RoR) 5.0.2.
- For the data mart, we employed the PostgreSQL 9.6 engine which conforms to the SQL standard. In addition, it includes a plugin for handling geographic data called PostGIS.
- For querying and developing dashboards, we employed the Metabase tool.

The final DW/BI architecture was defined as follows: the data sources are passed to the central repository, which is implemented in PostgreSQL, by means of the ETL processes developed in the Pentaho tool. Afterward, the information is presented to the users through a BI visualization tool called Metabase. The visualization is embedded in a web application developed with RoR.

F) Application

In this activity we describe all the applications developed for implementing the observatory.

1. *Admin RD:* This application was developed to manage disaster risk data. It allows for inserting, updating, showing, and deleting risk data for every neighborhood in the Santa Marta urban area. This application is based on the methodology proposed by the PNUD and the UNGRD [15], where all threat variables (intensity, frequency, and affected territory) and all vulnerability variables are scored with values 1, 2, and 3 (low, medium, and high, respectively).
2. *SM API:* As for climate data, we developed a Java application that carries out requests to the climate API to retrieve *json* files, processes them, and stores the relevant data in the relational database.
3. *Data mart querying application:*Given the number of variables, as well as the flexibility established in the business requirements, we decided to employ a tool that allowed us to carry out our own analyses through a friendly interface. This tool was also required to allow us to develop dashboards for solving specific requirements. Metabase allows information querying and offers several display forms to present them, from the most basic, which is the tabular format, to more complex forms, such as maps. In this point we once again employ the visualization framework in order to identify the most appropriate display form for each datatype. In order to develop the proposed dashboard, we evaluated their usability based on test sessions with the end-users.
4. *Disaster risk observatory for Santa Marta:* The web application for the Santa Marta disaster risk observatory was developed using RoR. It presents information related to risk knowledge (threat types, relevant articles) and embeds the dashboards developed in Metabase.

The disaster risk dashboard includes seven types of visualizations that facilitate disaster risk management.

The next activities are deployment, maintenance, and growth. These were not addressed in the study case, but they are related to the deployment of the system in a production setting as well as the improvements made to it after launch.

4. RESULTS

We carried out observatory evaluation based on a test that consisted of four business analysis cases proposed in the definition phase. With these cases we compared our tool and the traditional way to carry out the task, that is, using spreadsheet tools. The test-users had 10 minutes for carrying out the test for each case with each method. Next, we did a survey on the perception the test-users had on the tool. We evaluated both approaches with eight users. The analysis cases are described next:

- Analysis of the distribution of risk, threat, and vulnerability levels in a per-neighborhood basis for a given period.
- Identification of persons with reduced mobility in a neighborhood.
- Discovery of the income distribution of people for a given period.

Case 1: To analyze the distribution of risk, threat, and vulnerability levels in a per-neighborhood basis for a given period.

Case 2: To identify the neighborhoods where there is the highest number of persons exposed to an elevated risk level for a given period.

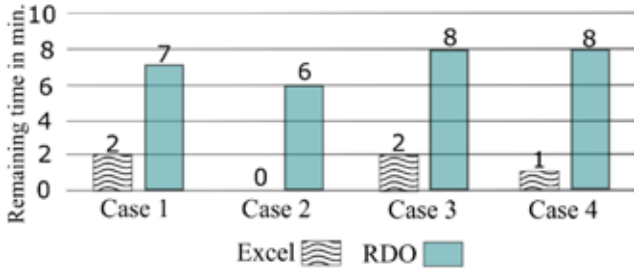
Case 3: To identify disabled persons in a neighborhood.

Case 4: To discover the income distribution of homes in a neighborhood.

The survey validated 4 groups of questions: user profile, usability, accessibility, and reliability.

In fig. 9 we depict the results of the tests. The cases are listed in the x-axis, while the y-axis shows the average remaining time for each case in minutes. By reviewing Fig. 5 we notice that there was less time left for each case when employing Excel and that in case 2, the activity was not finished. This shows that our tool is more effective for carrying out analyses.

Figure 5. A comparative plot of the cases solved correctly in Excel and our observatory.



5. CONCLUSIONS AND FUTURE WORK

Jointly applying Kimball's methodology and Munzner's framework allowed us to develop a strategy for creating flexible, fast risk observatories that fulfill end-user needs. Using Kimball's matrix makes it intuitive to learn the dimensional model, thanks in great part to the documentation provided by Kimball's group. Task abstraction through the verbs proposed by Munzner allows end-user requirements to be translated into a more technical language.

Carrying out research on existing tools for disaster risk management allowed us to choose the strategy that fitted the conditions that must be considered for developing disaster risk observatories.

The interactions proposed in the end-user application task guarantee success in fulfilling their tasks, as the prototypes are reviewed continuously, and any necessary adjustments are made. It should be noted that the visualizations presented are not definitive, there is a maintenance stage that allows the adaptation of points to adjust more to the needs of user groups, such as the number of graphics used in the dashboards or the creation of new ones. Data profiling and exploration tasks allow us to discover problems with the requirements quickly. One such case is not being able to solve a requirement as the data required for the task is not available. This reduces time and resources spent in unnecessary activities.

The proposed risk analyzes allowed intuitively to make decisions aimed at the process of disaster risk reduction. It is necessary to state that the analyzes that were made were of passive type, that is, they are analyzes to obtain reports, and measurements on the available data. It would be interesting and even pertinent to apply data mining techniques to perform types of active analysis.

Treating disaster risk management as a process enables us to start the project lifecycle earlier as we know what the starting point is, and it allows us to know which process will need to be developed afterwards.

As for future work, we must note that we will carry out the tasks of Deployment, Maintenance and Growth so that the Observatory can be used by many people and evolves to cover the following sub processes of Risk: Reduction and Disaster Management.

REFERENCES

[1] T. R. Malthus, *An Essay on the Principle of Population*. London: J. JOHNSON, 1798.

[2] Organización de las Naciones Unidas (ONU), "Más de la mitad de la población vive en áreas urbanas y seguirá creciendo," 2014. [Online]. Available: http://www.un.org/es/development/desa/news/population/world-urbanization-prospects-2014.html. [Accessed: 06-May-2017].

[3] S. Baas, S. Ramasamy, J. Dey de Pryck, and F. Battista, *Disaster risk management systems analysis A guide book*. Rome, 2008.

[4] Departamento Nacional de Planeación (DNP), "3.181 muertos y 12,3 millones de afectados: las cifras de desastres naturales entre 2006 y 2014," 2015. [Online]. Available: https://www.dnp.gov.co/Paginas/3-181-muertos,-21-594-emergencias-y-12,3-millones-de-afectados-las-cifras-de-los-desastres-naturales-entre-2006-y-2014-.aspx. [Accessed: 06-Jun-2017].

[5] Banco Mundial, "Análisis de la gestión del riesgo de desastres en Colombia: un aporte para la construcción de políticas públicas.," *Sist. Nac. Inf. para la Gestión del Riesgo Desastr.*, p. 438, 2012.

[6] O. D. Cardona, "Estimación holística del riesgo sísmico utilizando sistemas dinámicos complejos," 2001.

[7] O. D. Cardona, "INDICADORES DE RIESGO DE DESASTRE y Gestión de Riesgo de Desastre," 2007.

[8] S. Rahman, "Gestión del riesgo de desastres: Panorama general," *Banco Mundial*, 2016. [Online]. Available: http://www.bancomundial.org/es/topic/disasterriskmanagement/overview. [Accessed: 02-Feb-2017].

[9] UNISDR, "Marco de Acción de Hyogo para 2005-2015," *Conf. Mundical sobre la Reducción los Desastr.*, p. 25, 2005.

[10] Programa de las Naciones Unidas para el Desarrollo (PNUD), "Evaluación del Riesgo de Desastres," 2015.

[11] N. Angulo Marcial, "¿Qué son los observatorios y cuáles son sus funciones?," *Innovación Educ.*, vol. 9, no. 47, pp. 5–17, 2009.

[12] Estrategia Internacional para la Reducción de Desastres (UNISDR), "Terminología sobre Reducción del Riesgo de Desastres," Ginebra, Suiza, 2009.

[13] Congreso de la República de Colombia, "Ley 1523 de 2012." Bogotá D.C., 2012.

[14] E. Avendaño, "Estrategia para desarrollar Observatorio de Riesgo de Desastres integrando Business Intelligence y Visual Analytics," Universidad de los Andes, 2017.

[15] Programa de las Naciones Unidas para el Desarrollo (PNUD) and UNGRD, "Guía metodológica para la elaboración de Planes Departamentales para la Gestión del Riesgo." 2012.

[16] Inter-Agency Standing Committee (IASC), "INFORM In Depth," 2016. [Online]. Available: http://www.inform-index.org/InDepth. [Accessed: 20-Oct-2016].

[17] CAPRA, "CAPRA Probabilistic risk assesment program," 2017. [Online]. Available: http://www.ecapra.org/es. [Accessed: 01-Apr-2017].

[18] University of Melbourne, "Centre for Disaster Management and Public Safety," 2014. [Online]. Available: http://www.cdmps.org.au/. [Accessed: 18-Mar-2017].

[19] A. Rajabifard and J. von Dr Känel, "The Australia Disaster Management Program," 2013. [Online]. Available: http://admp.org.au/. [Accessed: 14-Mar-2017].

[20] Centre for Disaster Management and Public Safety, "An Intelligent Disaster Decision Support System (IDDSS)." [Online]. Available: http://www.cdmps.org.au/intelligent-disaster-decision-support-system-iddss/. [Accessed: 15-Mar-2017].

[21] THE UNIVERSITY OF MELBOURNE, "Risk Finder," 2016. [Online]. Available: http://apps.csdila.ie.unimelb.edu.au/riskfinder/. [Accessed: 14-Mar-2017].

[22] IDIGER, "Sistema de Información para la Gestión del Riesgo y Cambio Climático," 2016. [Online]. Available: http://www.sire.gov.co/. [Accessed: 12-Mar-2017].

[23] IDIGER, "Sistema Distrital - IDIGER," 2016. [Online]. Available: http://www.idiger.gov.co/web/guest/sistema-distrital. [Accessed: 27-Feb-2017].

[24] IDIGER, "Instituto Distrital de Gestión de Riesgos y Cambio Climático," 2016. [Online]. Available: http://www.idiger.gov.co. [Accessed: 13-Mar-2017].

[25] Servicio Geológico Colombiano, "Geoportal del Servicio Geológico Colombiano," 2014. [Online]. Available: http://geoportal.sgc.gov.co/geoportalsgc/catalog/main/home.page. [Accessed: 02-Feb-2017].

[26] G. G. Perez, "Teoría Social Del Riesgo Y Cartografía Aplicada a la Ciudad de Neuquén," *Boletín Geográfico*, vol. 32, no. 0326--1735, pp. 115–124, 2010.

[27] V. Guzman Mesa, D. Paez Barajas, A. RAJABIFARD, and M. Sanchez Puccini, "THE COLOMBIAN EMERGENCY RESPONSE PLATFORM (PCRE): DESIGN AND TESTING OF A SDI -," Universidad de los Andes, 2016.

[28] R. Kimball, M. Ross, W. Thornwaite, J. Mundy, and B. Becker, *The Data Warehouse Lifecycle Toolkit*, 2nd ed. Indianapolis; Canada: Wiley Publishing, Inc., 2009.

[29] T. Munzner, *Visualization Analysis and Design*, 1st ed. Boca Ratón, FL; London; New York, 2014.

[30] R. Kimball, M. Ross, W. Thornwaite, and B. Becker, *The Data Warehouse Toolkit*, 3rd ed. Wiley Publishing, Inc., 2008.

[31] G. R. Rivadera, "La metodología de Kimball para el diseño de almacenes de datos (Data warehouses)," Buenos Aires, 2010.

[32] L. A. Lozano, "Estrategia para Desarrollar Observatorios de Salud Integrando Inteligencia de Negocios y Analítica Visual," Universidad de los Andes, 2013.

[33] Gesellschaft für Technische Zusammenarbeit - GTZ, "Incorporar la gestión del riesgo en la planificación territorial. Orientaciones para el nivel municipal." p. 68, 2010.