Creation of a Data Warehouse using the F-Cube Factory Software to resolve problems with degrees of truth.

ZAPATA C. SANTIAGO, MARURI B. CHRISTIAN, ROJAS B. RONALD
Universidad Tecnológica Metropolitana
Departamento de Informática –Santiago- Chile
szapata@utem.cl, cmaruri@icci.cl, ronald.rojasb@gmail.com

Abstract
The following work shows a solution for high complexity problems, with the commissioning of a Data Warehouse System applied to historical databases, which delivers useful information for decision making support in environmental situations, oriented to resolve atmospheric pollutions problems generated by particulate material (PM10)

In this work, a hardware-software platform is deployed, which allows the use of OLAP cubes for the resolution of problems with big volume databases. In this case, the particulate material (PM10) of the city of Santiago, Chile, between the years 1997-2005, logged by the National Environment Commission (CONAMA)

For the following work, we use the software f-cube Factory, developed by Carlos Molina Fernandez PhD in his doctoral thesis, “imprecision and uncertainty in the multidimensional model: data mining applied” (DECSAI, Universidad de Granada, Span, 2005)

The software F-Cube Factory is an OLAP (Online Analytical Processing) system, which allows the definition, construction and query of data cubes (fuzzy and crisp). All realized from tables stored in a relational database. The reports generated by this software can be seen in table or graphic format, according with the researcher’s choice.

The first part of the works, describes the current situation of the Santiago’s basin, current predictive model used to decree environmental warnings and a multidimensional model proposition.

The second part consist in the creation of a Data Warehouse the can be used for the F-cube Factory tool to store the information relative to the atmospheric pollution in the city of Santiago de Chile, and so, get views and interesting dimensions generated from the stored data.

Keywords: Data Warehouse, F-Cube Factory, OLAP, Linguistic labels, Fuzzy Logic

1 Introduction
The pre-emergencies in Chile are established when the pollution levels exceeds the values indicated in the environmental law, this has a serious impact in the commerce, and in some critical cases it could become a sanitary emergency, because it’s dangerous to the people’s health, therefore it’s needed to count with methods capable to study the historical situation of the Santiago’s basin, which have characterized for having trouble with the pollution and the dust in suspension, and so to predict anomalous ventilation situations, that permits the authorities to act efficiently.

1.1 Current Prediction and measurement model
Currently, a predictive model created by Joseph Cassmassi [23] is used, the model was developed (starting from) from the air quality information measured by the Automatic Monitoring Network of air quality (MACAM II Network) and the tall metrological information from the central zone of the country.

The forecasting methodology of MP10 concentrations is based in calculus algorithms developed by applying statistical techniques of multiple regression variables, focused in find relations between possible predictor variables and a variable to predict. The possible predictors include observed weather variables, observed weather condition indexes, observed concentrations and expected variations in rates of emissions.

The Cassmassi model forecast the maximum value of average concentration in 24 hours of breathable particulates material (PM10), forecasted for 00-24 h period of the following day, expressed in (ug/m3), in each one of the stations of the MACAM 2 Network classified as PM10 Monitoring Stations with poblational representatively (EMRP). These are: Av. La Paz, La Florida, Las Condes, Parque O'Higgins, Pudahuel, Cerrillos and El Bosque, according with the resolution N°11481 of 1998 from SESMA.
The forecasted concentration for the next day is calculated by different equations for each air quality monitoring station. The required variables for the equation solving are obtained from the related information with the expected conditions shifting by day of the week, from the PM10 concentrations measured in the MACAMII Network, from tall meteorological information obtained from the radio probing realized by the Weather Direction of Chile and the weather conditions of synoptic and regional observed and forecasted scale for the region.

The operational application of this methodology considers two prediction algorithms for each monitoring station. A first algorithm includes the index of meteorological potential, forecasted for the next day. The second algorithm is based in observations only (same day and previous day). That way, if the first algorithm cannot be applied, the second one is used.

With the model previously described, environmental measures are enacted according with the following table:

<table>
<thead>
<tr>
<th>ICAP Level</th>
<th>Air Quality</th>
<th>Enacted Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-99</td>
<td>Good</td>
<td>None</td>
</tr>
<tr>
<td>100-199</td>
<td>Regular</td>
<td>None</td>
</tr>
<tr>
<td>200-299</td>
<td>Bad</td>
<td>Environmental Alert</td>
</tr>
<tr>
<td>300-399</td>
<td>Critical</td>
<td>Pre-emergency</td>
</tr>
<tr>
<td>400-499</td>
<td>Dangerous</td>
<td>Pre-emergency</td>
</tr>
<tr>
<td>500 or more</td>
<td>Exceed</td>
<td>Emergency</td>
</tr>
</tbody>
</table>

Table 1.1 ICAP Levels and Environmental Measures

All this, so the implemented system can become a decision-making support tool at the time of enacting environmental measures.

2 Multidimensional Model

For the following article we present a multidimensional model with explicit hierarchies. In first place, we define the structures of the dimensions and the hierarchies that we can define in them. An then we'll extend the definitions to represent uncertainty and imprecision [20][21].

Definition 1: A dimension is a tuple \(d = (l, \leq_d, l_\perp, l_T)\) where \(l = \{l_1, \ldots, l_n\}\) so that each \(l_i\) is a set of values \(l_i = \{c_{i1}, \ldots, c_{im}\}\) and \(l_i \cap l_j = \emptyset\) si \(i \neq j\), and \(\leq_d\) is a partial order relation between the elements of \(l\) so that \(l_i \leq_d l_k\) if \(\forall c_i \in l_i \Rightarrow \exists c_k \in l_k / c_i \subseteq c_k\). \(l_\perp\) and \(l_T\) are two elements of \(l\) so that \(\forall l_i \in l / l_i \leq_d l_\perp y l_i \leq_d l_T\).

We denote level to each element \(l_i\). To identify the level \(l\) of the dimension \(d\) we will use \(d.l\). One element of the dimension \(d\) and level \(l\) will be denominated \(c_{i_k}\). The special levels \(l_\perp\) and \(l_T\) will be called base level and top level respectively.

The domain of a dimension will be the set of all the values that appear in all the defined levels.

Definition 2: For each dimension \(d\), the domain is \(\text{dom}(d) = \bigcup l_i\).

The levels of a dimension are related by partial order. This relationship establishes hierarchy between them. A level \(l_i\) can have lower levels which are directly connected in the hierarchy. That is, levels lower than \(l_i\), there is no other lower level that \(l_i\) and that in turn greater than this. All the levels that meet this condition we will call for a level set of children's level

Definition 3: For the set \(H_i = \{l_i / l_i \neq l_k \land l_i \leq_d l_k \land \neg \exists l_i \leq_d l_k \leq_d l_i\}\)

We call it the set of children of the level \(l_i\).

With this definition, in any dimension, we would have that for each base level its set of children are a null set. At the same time, we would have levels immediately greater in the hierarchy; in this case, to set higher levels directly related to another is identified as parents set of the level.

Definition 4: Given \(l_i\)

\(P_i = \{l_i / l_i \neq l_k \land l_i \leq_d l_k \land \neg \exists l_k \leq_d l_k \leq_d l_i\}\)

We call it the set of parents of the level \(l_i\).

Definition 5: We considerate one set of attributes \(A_1, \ldots, A_n\) with domains \(D_1, \ldots, D_m\), we call fact to any \(h = (x_1, \ldots, x_m)\) such that \(x_i \in D_i\).
∀ \ i = 1, \ldots, n \ , \text{that is any n-tuple defined by the attributes' domains.}

Once we have the facts, to adjust the level of detail (granularity), these must be group by an aggregation operator, example: average, sum.

**Definition 6:** Let B(X) denote all the possible fuzzy bags defined using elements in X, and D_x be a numeric or natural domain, we define an aggregation operator G as a function G : B(D_x) \rightarrow D_x.

**Definition 7:** An object of type history is the recursive structure

\[ H^0 = \Omega \]
\[ H^{n+1} = (A, l_b, F, G, H^n) \]

Where

- \( \Omega \) is the recursive clause,
- \( F \) is the fact set,
- \( l_b \) is a set of levels \((l_{1b}, \ldots, l_{nb})\),
- \( A \) is an application from \( l_b \) to \( F(A : l_b \rightarrow F) \),
- \( G \) is an aggregation operator.

This structure enables details levels of the Data Cube to be stored while it is operated on so that it may be restored to a previous level of granularity.

**Definition 8:** A Data Cube is a tuple \( C = (D, l_b, F, H, A) \) such that \( D = (d_1, \ldots, d_n) \) is a set of dimensions, \( l_b = (l_{1b}, \ldots, l_{nb}) \) is a set of levels such that \( l_{ib} \) belongs to \( d_i \), \( F = R \cup \emptyset \) where \( R \) is the set of facts and \( \emptyset \) is a special symbol. \( H \) is an object of type history and \( A \) is an application defined as \( A : l_{1b} \times \ldots \times l_{nb} \rightarrow F \), giving the relation between the dimensions and the facts defined.

**Definition 9:** We say a Data Cube is basic if \( l_b = (l_{1b}, \ldots, l_{nb}) \) and \( H = \Omega \).

**Definition 10:** For each value \( \bar{c} = (\bar{c}_1, \ldots, \bar{c}_j, \ldots, \bar{c}_n) \) belonging to \( l_r \), we have the set

\[ F_{\bar{c}} = \bigcup \{ F_{\bar{c}} / c_{kp} \in l_k \wedge c_{kp} \subseteq c_{ij} \text{ si } l_r \neq l_b \}
\[ \{ h / h \in F \wedge \exists \bar{c}A(\bar{c}) = h \} \text{ si } l_r = l_b \]

Once we have the definition of \( F_{\bar{c}} \), we can introduced de roll-up operation.

Other definitions are in [20][21].

### 3 Data Source

The data used were provided by the Environment National Commission of Chile (CONAMA), who obtains their data from its monitoring network called MACAM II (automatic monitoring of air pollutants) which has 8 stations in Santiago. These data are from 1997 to 2005 and are stored in csv format (comma separated value) separately.

These are the main pollutants in The Santiago’s Basin:

- **Breathable Particulate Material:** Depending on their aerodynamic diameter (10 and 2.5 microns) is in the air as dust, smoke or other aerosols. The most important emission sources are vehicular emissions (diesel), power plants that use fuels and construction activities. The highest concentrations occur during the winter months. The prolonged exposure to this pollutant provokes an increased frequency of lung cancer, premature death, severe respiratory symptoms and irritation of eyes and nose.

- **Carbon Monoxide (CO):** It is created from the incomplete combustion of carbon compounds and some industrial and biological processes. The main source comes from the vehicles emissions. It affects health by interfering with the transport of oxygen to the heart, the brain and other muscles.

- **Sulfur Dioxide (SO2):** Comes from burning fossil fuels, smelting of ores containing sulfur, and other industrial processes. It is emitted
by diesel vehicles. This gas reacts with a wide variety of aerosols, so that its action is enhanced by the presence of particulate material. This substance has irritant effects on the airways, resulting in an obstructive bronchitis and bronchial obstruction. In addition, their synergistic effect with other pollutants can be highly aggressive.

**Ozone (O3):** This secondary pollutant is the main component of photochemical smog, and one of the strongest oxidizing agents, formed from the action of sunlight in an indirect way in nitrogen oxides and volatile organic compounds in the troposphere, and the action of the same in the molecules of ozone in the stratosphere. The formation of ozone can occur on time scales of hours to several days, as a result, ozone concentrations are temporary, spatially separated from the precursor sources and their concentrations in the atmosphere are strongly dependent on the weather. The symptoms that have been reported for this pollutant include demonstration of cough and headache, sore eyes, nose and throat, chest pain, increased mucus, wheezing, airway closure, lethargy, malaise and nausea, and increased in the incidence of asthma attacks.

### 3.1 MACAM network

The MACAM II network is composed by 8 stations en Santiago; such are shown in the Figure 3.2:

![Figure 3.2 MACAM Network Stations](image)

Besides of these eight stations, three stations were added later (Cerro Navia, Cerrillos2 and La Dehesa), for to have a greater amount of data available and, therefore, provide greater reliability to data.

### 4 Data Warehouse

The data described in the previous chapter are extracted, transformed and loaded with the ETL software Pentaho Data Integration. A group of comma-separated files (CSV) generates one unique table with all the data. Besides eliminating those there was no measure (without value or -99), as shown in table 4.1:

<table>
<thead>
<tr>
<th>Date</th>
<th>Hour</th>
<th>GAS Station</th>
<th>Measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td>YYYYMM DD</td>
<td>HH</td>
<td>GAS_1</td>
<td>Statio n1</td>
</tr>
<tr>
<td>YYYYMM DD</td>
<td>HH</td>
<td>GAS_1</td>
<td>Statio n2</td>
</tr>
<tr>
<td>YYYYMM DD</td>
<td>HH</td>
<td>GAS_2</td>
<td>Statio n3</td>
</tr>
<tr>
<td>YYYYMM DD</td>
<td>HH</td>
<td>GAS_2</td>
<td>Statio n4</td>
</tr>
</tbody>
</table>

**Table 4.1 Data formatted to use with F-Cube Factory**

### 4.1 Data flow elements

![Figure 4.1 Data flow in ETL Pentaho Data Integration](image)

The data flow inside the ETL includes the following stages:

- **Load CSV File**: Input operation, this loads the initial file(s) to work, each file contain the data of one gas measured in all stations.

- **Row Normaliser**: Transformation Operation, the ETL moves all the stations into one column named ESTACIÓN and all the measures into one column named MEDICION.

- **Filter Rows**: Flow Operation. Once the records are normalized, proceed to eliminate records those that do not report
measurement. (In the CSV files, the measure displays the value -99).

**Generate Rows**: Input operation. This generates constant values to the table, in this case, each gas will receive an ID that must be added as column in each normalized row. This is with the purpose to identify which gas belongs the record.

**Join rows (Cartesian Product)**: Union operation, this operation allows joining the rows generated in the previous operation, with the rows previously filtered and normalized from the CSV file.

**Table output**: This output operation, it gets the seven Cartesian product (one for each gas) and combine them in a database table, in this case in a MySQL database table.

### 4.2 Creating the Cube

<table>
<thead>
<tr>
<th>Fact</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEDICION</td>
<td>Magnitude of measurement of gases and particulate material.</td>
</tr>
<tr>
<td>ICAP</td>
<td>Index used to categorize the amount of pollutants (view Table 1.1)</td>
</tr>
</tbody>
</table>

**Table 4.2 Sesma Cube Facts**

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>FECHA</td>
<td>Date of the sampling, in format year, month, day (YYYYMMDD), is an integer value.</td>
</tr>
<tr>
<td>ANIO</td>
<td>Year of the sampling, format YYYY, is an integer value between 1997 and 2005.</td>
</tr>
<tr>
<td>MES</td>
<td>Month of the sampling, format MM, it's an integer value between 1 and 12.</td>
</tr>
<tr>
<td>DIA_ANIO</td>
<td>Day of the year of the sampling, format DDD, it's an integer value between 1 and 366.</td>
</tr>
<tr>
<td>DIA_MES</td>
<td>Day of the month of the sampling, format DD, it's an integer value between 1 and 31.</td>
</tr>
<tr>
<td>DIA_SEMANA</td>
<td>Day of the week of the sampling; the values are Monday, Tuesday, Wednesday, Thursday, Friday, Saturday and Sunday.</td>
</tr>
<tr>
<td>HORA</td>
<td>Hour of the sampling, format HH, it's an integer value between 0 and 23.</td>
</tr>
</tbody>
</table>

**TEMPORADA**

Season of the year of the sampling, its values are:
- SPRING (September 21- December 20),
- SUMMER (December 21 – March 20),
- FALL (March 21 - June 20),
- WINTER (June 21 - September 20), it’s a varchar value

**ESTACION**

Station, place of the sampling. Varchar value. (Figure 3.2)

**GAS**

Name of measured gas in the sampling (detailed on chapter 3)

**ESTADO**

Corresponds to the air quality level, this dimension is only used with the gas MP1024h. Possible values: BUENO, MALO, REGULAR, CRITICO, PELIGROSO y EXCEDE (detailed on Table 1.1)

**Table 4.3 Sesma Cube Dimensions**

### 5 Architecture of fuzzy queries

We want to know if there is any correlation between the ICAP index and the temperatures registered during the year.

We consider the multidimensional model shown in Figure 5.1. There are 612976 records and a fact of float data type, named ICAP.

**Figure 5.1 Crisp cube loaded in F-Cube factory**

On this multidimensional model, we want to add the temperature level, to classify the months considering them cold, warm and hot.
If we ask one people for this relation, the answer usually is express in linguistics labels (as show in table 5.1).

<table>
<thead>
<tr>
<th>Month</th>
<th>HOT</th>
<th>WARM</th>
<th>COLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>Very high</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>February</td>
<td>Very high</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>March</td>
<td>Medium</td>
<td>Medium</td>
<td>Medium</td>
</tr>
<tr>
<td>April</td>
<td>Medium</td>
<td>Medium</td>
<td>Medium</td>
</tr>
<tr>
<td>May</td>
<td>None</td>
<td>Medium</td>
<td>High</td>
</tr>
<tr>
<td>June</td>
<td>None</td>
<td>None</td>
<td>Very high</td>
</tr>
<tr>
<td>July</td>
<td>None</td>
<td>None</td>
<td>Very High</td>
</tr>
<tr>
<td>August</td>
<td>None</td>
<td>Medium</td>
<td>Medium</td>
</tr>
<tr>
<td>September</td>
<td>None</td>
<td>Very High</td>
<td>None</td>
</tr>
<tr>
<td>October</td>
<td>None</td>
<td>Very High</td>
<td>None</td>
</tr>
<tr>
<td>November</td>
<td>Medium</td>
<td>Very High</td>
<td>None</td>
</tr>
<tr>
<td>December</td>
<td>Very High</td>
<td>None</td>
<td>None</td>
</tr>
</tbody>
</table>

Table 5.1 linguistics labels for temperatures and months of the year.

F-Cube Factory can it to be used for define this relation using the information that has been presented. Figure 5.2 and 5.3 shows the creation of one fuzzy cube from a discrete cube.

To add the relation between the months and the temperature, just we have that select what values are grouped for each category of temperature and define the membership degrees using the linguistics labels of the fig, 5.5, Figure 5.6 and 5.7 shows the final relations for the temperature levels and months.

After, we can use this new level in the dimension for analyze data. A question for this analysis will be ask for the average level of pollutants (measured by the variable ICAP) according to the temperature of month. To get the result, we have to select the dimension FECHA (slice operation) and the Temperature level (roll-up) as shown in figure 5.11.
Figure 5.8 Creation of fuzzy query cube.

Figure 5.9 Dimension selection
In this case, the last step is to select the facts of the resulting data cube, as shown in fig. 5.10. According to the query, we have to select an aggregation operator of ICAP index (Average). Also, we define the constraints if it’s required, as shown in figure 5.11.

Figure 5.10 Facts selection

Figure 5.11 Constraints selection
After running the query, we have a new dataset with the results, as shown figures 5.14 and 5.15.

Figure 5.12 Calculated query cube
As that in the query of cube crisp, we can access to the data either by the sample of the information registered, see fig 5.16 and 5.17 or by the use of graphics, see fig 5.18 and 5.19.

Figure 5.13 Query cube information page.
6 Conclusion and future work

Cube factory is an interesting tool to convert bidimensional tables into multidimensional olap cubes. Most cases faster than a direct mysql query. The software is suitable to run in workstations and personal computers, although the system requirements and computing time rise in a similar proportion as the size of the data source.

The lack of documentation, tech support, software updates and source code access, makes difficult to the user community develop new versions and patches. That puts F-Cube Factory in disadvantage against other solutions in the market (Open Source and commercial). However the software can be a valid alternative to use in teaching subjects related with OLAP, fuzzy logic and data mining.

By the fuzzy data model, we can obtain information hard to extract with the crisp (traditional) logic (i.e.: levels of relations with degrees of truth). In turn, the obtained results in the fuzzy way fits to historical reality of the Santiago’s basin. However this doesn’t permit us to predict future behaviors.

The model also, avoids the need to regenerate the cube from the beginning when is needed a higher detail in the previously done queries on a subset of a cube.

In general, the predictive models must be in constant update and upgrading, due to the temporal or permanent change of the environment conditions of the Santiago’s basin. An important element to increase the accuracy of such models is the data source. While more data is delivered for analysis (i.e.: temperature, fallen water, moisture, wind speed), the greater the fiability of the information generated for support taking decision.

Although the current models comply the law, they have been criticized for their low accuracy, which has generated false positives (environmental measures enacted when the conditions don’t need it), also occur episodes that not environmental measure is enacted, when the conditions of ventilation were required it.
Bibliography