Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

Design and implementation of a climatic data warehouse

J. Velázquez-Álvarez[1] & J. Torres-Jiménez[2]
[1]*Hydrologic Technology Department, Mexican Institute of Water Technology, México*
[2]*Computer Science Department, ITESM Campus Morelos, México*

Abstract

The design and implementation of a climatic data warehouse is described. Data is daily measures of nine climatic variables which include temperature, rainfall and evaporation, among others; this data is recorded at weather stations installed in different locations of Mexico. The data warehouse design is based on the dimensional model which allows for high performance and yields data diagrams that are easy to understand and intuitive to final users. Also some aggregates were built in order to make more efficient the query processing. The data warehouse shown also provides geographic information system capabilities that allow the user to make spatial queries. The described system is useful to analyze and extract data that is required in climate studies and other disciplines.

1 Introduction

Climate studies have become more important over the last years mainly due to the extreme climatic events (droughts and floods) happened in many parts of the world, and particularly in Mexico.

The Mexico National Weather Service (SMN) is the institution in charge of management all data related to weather forecast and climate in general. The SMN has many software systems that provide the users with different kind of information. One set of these systems provides historical daily measures of 9 climatic variables: temperature (max., min. and value at 8 a.m.), accumulated rainfall and evaporation, sky status and presence of electric storm, fog and hail. This data is recorded at a 5,575 weather station national network. Some weather stations contain data recorded from 1893. One of the problems found in these systems is that they are not flexible enough to answer the unforeseen and great amount of queries that users submit, and report data in the wide variety of formats that final users request. There are also representation problems (some systems uses 0 and 1 to indicate electric storm in one day, and others use False or True). Another problem is related to consistency: for a station and a date given, sometimes the reported value is not the same.

Because of this, a climatic data warehouse was proposed to solve the problems above mentioned. Also, geographic information system capabilities were added to this system in order to visualize maps and the location of weather stations selected by a logical condition.

2 Data warehousing concepts

2.1 Data warehouse definition and components

A data warehouse is an historical database that is the result of integrating data from many source systems. The data warehouse design allows querying and analyzing huge amounts of data in order to provide information to decision making in one organization. Information that is integrated and put in a Dimensional Data Warehouse (DDW) often comes from traditional OLTP (On Line Transaction Processing) systems. One difference between OLTP and DDW systems is that the first ones seek to make efficient the record and management of all transactions that happen within an organization. On the contrary, a data warehouse usually is not on line and emphasis is put on extracting relevant information about data coming from OLTP systems. Other differences are related to modeling technique, horizon over time and tools used to information extraction [1], [2].

Figure 1 shows the main three components of a data warehouse. The first one is the adquisition component which includes tools for extracting data from the different data sources (operational databases and external sources such as text files); for cleaning, transforming and integrating this data; for loading data into the data warehouse; and for periodically refreshing the warehouse to reflect updates at the sources. The second component (storage) is a large, physical database that holds a large amount of information integrated from a wide variety of sources. This is the data warehouse itself. The physical equipment where the database is stored, is called the presentation server, [3]. The third component includes all the different applications that access and make use of the information stored in the warehouse. These are tipically OLAP (On Line Analytical Processing), Data mining and Decision support systems (DSS) applications. Finally, there is a repository to store and manage metadata, and tools for monitoring and administering the warehousing system.

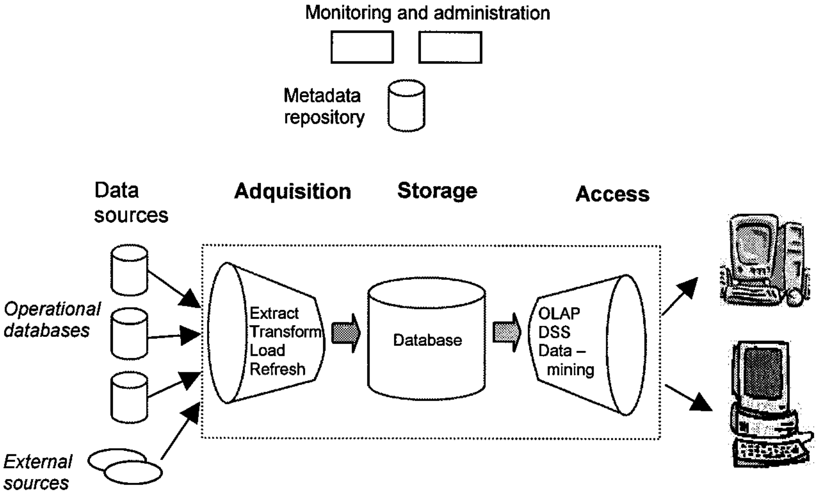

Figure 1: The main components of a data warehouse

2.2 The dimensional model

Dimensional modeling is a logical design technique that seeks to present data in a standard framework that is intuitive and allows for high – performance access. This technique seeks to visualize a database as a “cube” of three or more dimensions (hyper-cube) which allows slicing and dicing along each of its dimensions. In a relational environment, every dimensional model is composed of one table with a composite primary key, called the *fact table*, and a set of smaller tables called *dimension tables*. Each dimension table has a single primary key that correspond to one of the components of the composite key in the fact table. This characteristic star-like structure is often called a *star join schema.*

The fact table is where the numerical measurements of the business are stored. The detail with which the measurement is made is called the *grain* of the fact table. Facts can be classified as additive, semi-additive and non-additive. Additive facts are those that can be added along any dimension; dollars, units and costs are additive because it makes sense to add any of them across every combination of time, product or store. Semi-additive facts can be added along only some of the dimensions. Inventory levels and temperature are examples of semi-additive facts (they can be added just across time to get an average value). Non-additive facts simply can’t be added at all.

A table dimension is one of a set of companion tables to a fact table. A dimension is a collection of text-like attributes that are highly correlated with each other. Each dimension is defined by its primary key that serves as the basis for referential integrity with any given fact table to which it is joined. The textual attributes (fields) are the basis for constraining and grouping within data warehouse queries and they are equal candidates to become row headers in a SQL query.

Two important operations that can be made in a data warehouse is drilling down and drilling up. Drilling down means obtaining more detail by adding a row header to an existing SQL request. Drilling up is the opposite operation. Kimball et. al. [3] show that the dimensional model is the only feasible solution to implement efficient data warehouses.

3 Design of the climatic data warehouse

3.1 Climatic data sources considered

Currently, the Mexico National Weather Service (SMN) has 5 software systems that are used in capturing and querying surface daily data that is measured in weather stations. These systems were considered as data sources for implementing the data warehouse. The following paragraphs give a short description of them.

SIC and DAT322. The climatic information system (SIC) contains information that is recorded in 5,575 weather stations, the period of time with data goes from 1921 to 1990. This system was developed in Oracle, on a Unix environment. DAT322 runs on Windows environments, holds data of 322 stations (from 1893 to 1998) and was developed in Visual FoxPro. Both systems provide some geographic information capabilities: the first one uses Arc/Info software to this purpose; the second one uses an ActiveX component called Map Objects.

The entity – relationship model is the same for both systems. In this logical model the daily data are stored in a table where one record is made up by the weather station key, the date (day / month / year) and the nine climatic variables measured in that site and on that date. This design provides great flexibility to extract data and compute statistics in any interval of time either it be a year, a month or an interval defined by two dates. One problem found in these systems is that there aren't tools to extract data in several formats. Also, there is a performance penalty caused by snowflaking in the table that holds attributes for the weather station entity. Snowflaking is the removal of low cardinality fields to separate tables and linked back into the original table with artificial keys [3].

CLICOM, SICLIM and ERIC. CLICOM was developed by the World Meteorological Organization in the early 80's, uses Data Ease as database management system and runs on DOS environment. Because CLICOM was created by that time, it lacks a user-friendly interface; its main strength is the good quality control it enforces to get data into the system. SICLIM is a system that runs on Windows environment, was developed in Visual Basic and provides geographic lookup capabilities through using an ActiveX component (Map Objects), just as DAT322 system. The database is managed with Access and the period of time with data is from 1921 to 1990. In this system, floating point variables such as rainfall, temperature and evaporation are multiplied by 10 and stored as 2-byte integers, which reduces the space required in secondary memory. Thus, a rainfall value of 17.5 is stored as 175.

ERIC is a software system that also runs on Windows and holds data from 1921 to 1998. It was developed in C language. Its main strength is the fast access to the information as it uses propietary binary data structures to store data. The storage used by this system is the smallest because sky status and boolean variables (hail, storm and fog) are stored in two bits (0 - it didn't happen, 1 – it happened, 3 – missing data). Also, floating point variables are stored as integer numbers. This system doesn't have geographic information tools, and the number of analysis and query tools is reduced.

The Entity Relationship model for these 3 systems is shown in figure 2. In this design each climatic variable is stored in a separated table where one record is made up by the station key, month, year and the 31 values measured in that month. One problem with this design is that obtaining statistics for a period of time different to one month, implies more programming effort. Also, there is a snowflaking problem with the weather station table, as mentioned in the paragraph that describes SIC and DAT322 systems. This design doesn't give enough possibilities to get data out in the many different formats final users require.

3.2 Dimensional design of the climatic data warehouse

The design was made by following the four step process proposed by Kimball [2], which consists in defining the following four issues:

Business Process to Model. The business process to model is the query and analysis system of surface data measured in the national network of non-automatic weather stations. The legacy systems that support this process are five and are those described in section 3.1

Grain of the Business Process. The grain is a daily detail. This is the minimum time resolution with which the climatic variables are recorded.

Dimensions. The dimensions are three:

Time.- The attributes for this dimension are time key, day number in month, week number in year, month, season of the year and year.

Weather station.- This dimension contains attributes related to the location of the weather station, characteristics of the data measured on the station, and station type. Location attributes are latitude, longitude and altitude, among others; attributes that describe data measured at the weather station are first and last date in which each variable was recorded, missing data (given as a percentage), and number of years with data.

Climatic Variable. This dimension provides information about each climatic variable. Attributes that make up the dimension are variable key, name, description, measurement units, type and abbreviation.

Fact table. A fact in this process to model is the value of a variable measured at one station and on a date given. So, the attributes of this table are weather station key, time key, variable key, and measured value Unlike the designs described at section 3.1, one record in this fact table will hold the value of one

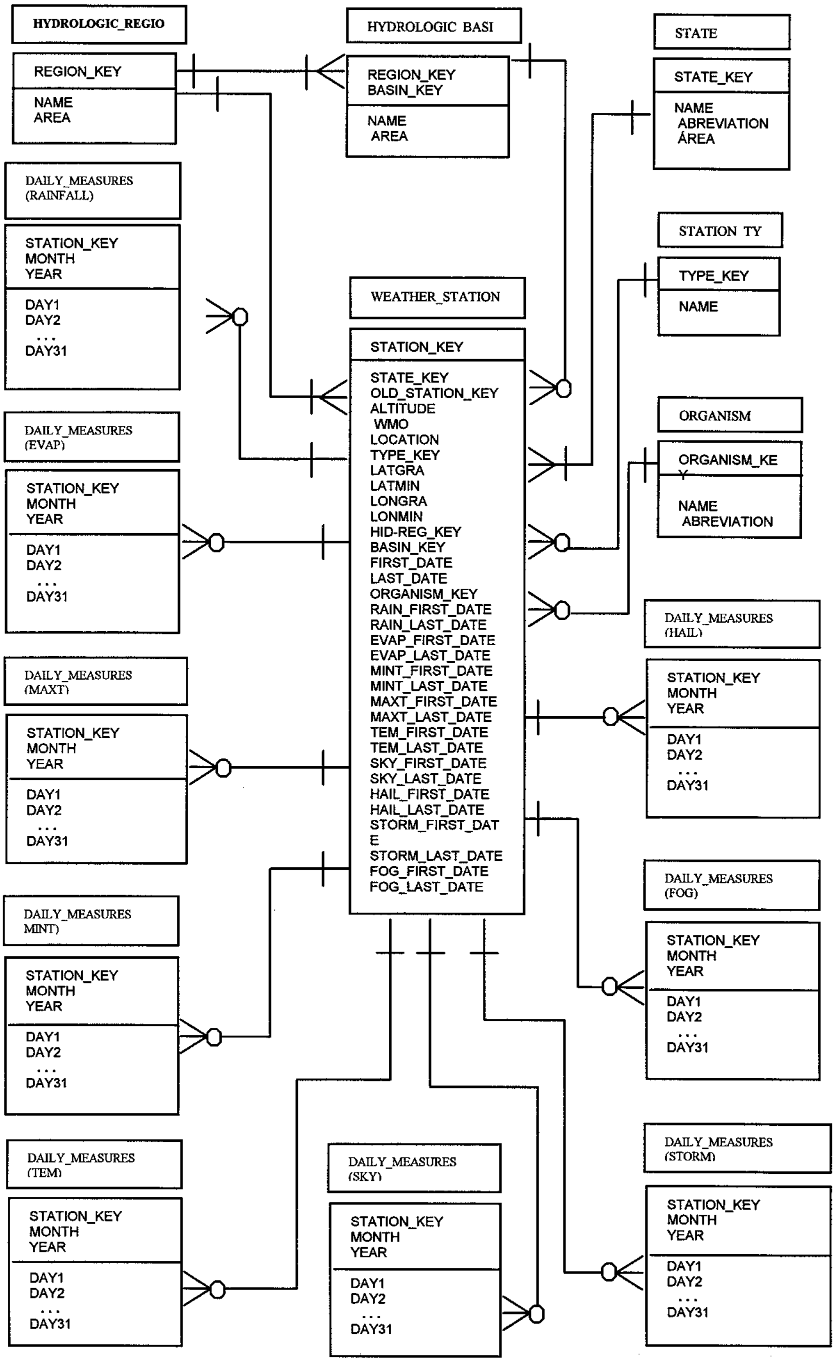

Figure 2: The Entity – Relationship model of the CLICOM, SICLIM and ERIC Systems

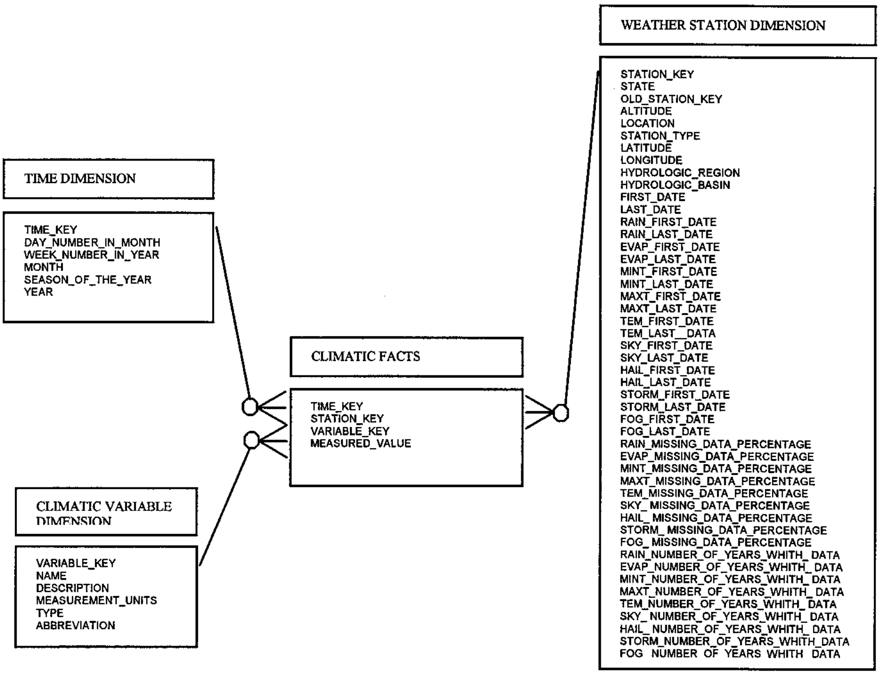
WEATHER STATION DIMENSION
STATION_KEY
STATE
OLD_STATION_KEY
ALTITUDE
LOCATION
STATION_TYPE
LATITUDE
LONGITUDE
HYDROLOGIC_REGION
HYDROLOGIC_BASIN
FIRST_DATE
LAST_DATE
RAIN_FIRST_DATE
RAIN_LAST_DATE
EVAP_FIRST_DATE
EVAP_LAST_DATE
MINT_FIRST_DATE
MINT_LAST_DATE
MAXT_FIRST_DATE
MAXT_LAST_DATE
TEM_FIRST_DATE
TEM_LAST__DATA
SKY_FIRST_DATE
SKY_LAST_DATE
HAIL_FIRST_DATE
HAIL_LAST_DATE
STORM_FIRST_DATE
STORM_LAST_DATE
FOG_FIRST_DATE
FOG_LAST_DATE
RAIN_MISSING_DATA_PERCENTAGE
EVAP_MISSING_DATA_PERCENTAGE
MINT_MISSING_DATA_PERCENTAGE
MAXT_MISSING_DATA_PERCENTAGE
TEM_MISSING_DATA_PERCENTAGE
SKY_ MISSING_DATA_PERCENTAGE
HAIL_ MISSING_DATA_PERCENTAGE
STORM_ MISSING_DATA_PERCENTAGE
FOG_ MISSING_DATA_PERCENTAGE
RAIN_NUMBER_OF_YEARS_WHITH_ DATA
EVAP_NUMBER_OF_YEARS_WHITH_ DATA
MINT_NUMBER_OF_YEARS_WHITH_ DATA
MAXT_NUMBER_OF_YEARS_WHITH_ DATA
TEM_NUMBER_OF_YEARS_WHITH_ DATA
SKY_ NUMBER_OF_YEARS_WHITH_ DATA
HAIL_ NUMBER_OF_YEARS_WHITH_ DATA
STORM_NUMBER_OF_YEARS_WHITH_DATA
FOG NUMBER OF YEARS WHITH DATA
TIME DIMENSION
TIME_KEY
DAY_NUMBER_IN_MONTH
WEEK_NUMBER_IN_YEAR
MONTH
SEASON_OF_THE_YEAR
YEAR
CLIMATIC FACTS
TIME_KEY
STATION_KEY
VARIABLE_KEY
MEASURED_VALUE
CLIMATIC VARIABLE DIMENSION
VARIABLE_KEY
NAME
DESCRIPTION
MEASUREMENT_UNITS
TYPE
ABBREVIATION

Figure 3: Dimensional model of the climatic data

variable. This design has many advantages: the possibility of adding new variables in the future, and a great flexibility to query data, compute statistics and extract data in many formats.

The dimensional model for the climatic data is shown in figure 3. In this model, the fact table is located at the center and dimensions tables around it. Dimension tables are joined to the fact table by their primary key.

3.3 Design of aggregates

Since the main goal of a data warehouse is to make efficient the query processing, often its physical design considers using aggregates and special indices. Aggregates are pre-computed and pre-stored summary data that are used to speed up queries. As to indices, bit map, join and multikey join indices are commonly used in decision support systems [4], [5].

In this work, aggregates were designed by following the practical approach proposed by Kimball et. al. [3]. Firstly, an analysis was made to determine the most common querys involving summary data, the attributes used in them and combinations of attributes. It was detected that many queries that requested summary data involved mainly attributes of the time dimension. All the possible combinations of these attributes and the size of each (given as number of records) were analyzed to choose aggregates that offered the highest benefits in answering a query. In this part, some heuristics proposed in [6], [7] were used. It was concluded that the aggregate that contained summary data, grouped by month and year, was the most useful.

Aggregates were designed also as star join squemas, with an aggregates fact table at the center and reduced dimension tables surrounding it. A dimension table get reduced when some attributes are eliminated as a result of making an aggregation. For example, once data are summed by month and year, the day, week_of_the_year and season_of_the_year attributes, in the time dimension, are no longer needed, and they are eliminated from the time dimension. Three aggregates fact tables were considered: one for continuous variables (rainfall, evaporation and temperature), one for discrete variables (electric storm, hail and fog), and one for the sky status variable. Dimension tables are shared by all the three fact tables. Each aggregates fact table has also time_key, weather_station_key and variable_key attributes.

4 Implementation

The data warehouse prototype was implemented considering just 227 weather stations, because there was not enough disk space to store data from all the stations. These weather stations were the ones that had the more complete set of data. The general period of time with data was from 1893 to 1999.

To get data out from the legacy systems several tools were developed, in fact one for each system due to the wide variety of database systems and formats. Computer programs were made mainly in C language. The data warehouse prototype was made by using Access as RDBMS system, and Visual Basic to create the graphic user interface and access data. Programs to graphic results were made in Visual C++.

During the extraction of data from the source systems an analysis was carried out in order to detect errors and inconsistencies. This analysis included checking values were between valid ranges, and obtaining maximum and minimum values, period of time with data, and missing percentage. Data cleaning was made by comparing data and choosing the values that were on valid ranges. In those cases where it was impossible to find out the right values, the data was discarded. A report with the errors found was delivered to the SMN. Also, at this stage of the development, naming and representation problems were solved. Later, data coming from the several sources was merged, transformed and put on text files with the format required by the designs described on sections 3.2 and 3.3. Finally, data was loaded by using the import functions that Access offers.

Geographic lookup capabilities were implemented with Map Objects software. Maps include states, hydrologic regions and basins, and location of weather stations. Also, functions related to maps display were included: zoom in, zoom out, pan and identify.

5 Results

Figure 4 shows the main dialog window of the data warehouse prototype; it displays a map of Mexico and location of weather stations (depicted as points). White points represent the 227 weather stations that were chosen to implement the data warehouse; black points depict the 5,575 weather stations that make up the national network. The prototype allows to navigate through the attributes of the weather station dimension and elaborate a logical condition in order to select a set of stations that meet such a condition. Thus, it is possible to submit a query like this: " State = 'Morelos' AND Number_of_years_with_rain_data > 20 AND rain_missing_data_percentage < 10 ". Selected stations are highlighted on the map. Once a set of stations has been selected, the query and analysis tools that the system offers are enabled.

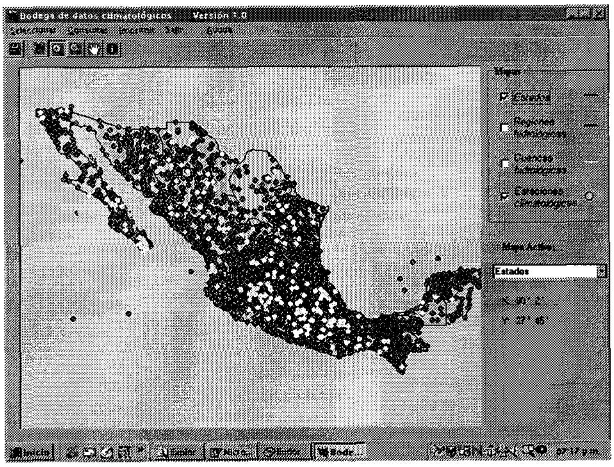

Figure 4: Location of weather stations in Mexico

The data warehouse includes tools to extract daily and summary data of the 9 climatic variables, as well as statistics (maximum, minimum and average values). Statistics can be obtained for one station or for a set of stations located within a geographic area, and grouped by any of the attributes of the time dimension.

The system also allows to compute and display the daily average values of a climatic variable, computed over a 30 - year period of time, at one location. These values are called *the climatic normal.* Figure 5 shows the climatic normal of rainfall for one weather station.This information is useful to know the average distribution of a variable over the year, in one location, and estimate how dry, normal or humid a given year has been when comparing to the average distribution of rainfall at that location.

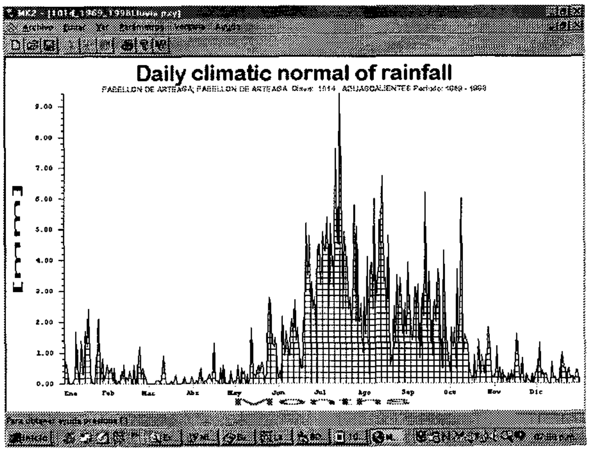
MX2 - [1014_1969_1998Lluvia.pxy]
Archivo Editar Ver Parámetros Ventana Ayuda
Daily climatic normal of rainfall
PABELLON DE ARTEAGA; PABELLON DE ARTEAGA Clave: 1014 AGUASCALIENTES Periodo: 1969 - 1998
9.00
8.00
7.00
6.00
5.00
4.00
3.00
2.00
1.00
0.00
[mm]
Ene
Feb
Mar
Abr
May
Jun
Jul
Ago
Sep
Oct
Nov
Dic
Para obtener ayuda presione F1
Inicio

Figure 5: Daily climatic normal of rainfall at one weather station

6 Conclusions

The proposed data warehouse prototype solves problems found in the software systems used currently at the Mexico National Weather Service to manage daily climatic data. Firstly, it makes a data integration with which naming, representation and consistency problems are removed.

Secondly, the design based on the dimensional model allows adding new variables without significant changes to report writers and query tools, yields a data diagram that is very simple and intuitive to the final users, and gives too much flexibility to accept unexpected queries and to report data in the many formats required by the users. Also, this design allows for high performance to access the information.

References

[1] Kimball, R., *The Data Warehouse Toolkit*, John Wiley & Sons, Inc., 1996.

[2] Torres-Jiménez J., *Notes of the course on Decisions Support Systems*. ITESM Campus Morelos. Mexico, 2000.

[3] Kimball, R., Reeves, L., Ross, M. & Thornthwaite, W., *The Data Warehouse Lifecycle Toolkit*. John Wiley & Sons, Inc., 1998.

[4] Chaudhuri, S. & Dayal, U., An overview of Data Warehousing and OLAP Technology. *ACM SIGMOD*, Record 26 (1), March, 1997.

[5] Sarawagi, S., Indexing OLAP Data. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, pp. 37-44, 1997.

[6] Harinarayan, V., Rajaraman, A. & Ullman, J.D., Implementing data cubes efficiently. *Proc. of the ACM SIGMOD Conference on Management of Data*, pp. 205-216, 1996.

[7] Baralis, E., Paraboschi, S., & Teniente, E., Materialized view selection in a multidimensional database. *Proc. of 23rd Int. Conf. On Very Large Databases*, Morgan-Kaufmann, pp. 25-29, 1997.