A Data Warehouse Architecture for MeteoSwiss: An Experience Report

Christian Häberli
MeteoSwiss
Krähbühlstr. 58
8044 Zurich, Switzerland
chi@sma.ch

Dimitrios Tombros
Swiss Technology Consulting Group AG
Technoparkstr. 1
8005 Zurich, Switzerland
dimitrios.tombros@stcg.ch

Abstract

In this experience report we consider some issues in the development of a data warehouse for meteorological and climatology data at MeteoSwiss. The paper describes the meteorological and climatology data process, the data warehouse architecture developed to support this process while integrating legacy systems and some of the modelling issues encountered.

1 Introduction

Efficient data storage and manipulation is a prerequisite in the meteorological and climatology domain. Large quantities of numerical and multi medial data are collected on a frequent basis from automatic measurement stations. This data includes measurements of several weather parameters on a 10-minute, hourly or longer basis. The collected data undergoes various quality control and consistency check procedures during its lifetime that effectively almost permanently update it. At the same time the data is continuously used for analysis by meteorological and climatology applications. It often undergoes aggregations according to special formulas. Thus different kinds of use patterns must be supported which pose severe performance problems.

In this paper we describe the data warehouse architecture developed for the Swiss national weather service, MeteoSwiss. We describe in section 2 how our work compares to other existing systems in this and related application domains. In section 3 we present the data life cycle that the system has to support and in section 4 we describe the developed architecture. In section 5 we discuss some data modeling issues.

2 Contributions and Related Work

The application of data warehouse technology in the domain of scientific data can be advantageous for manipulating large quantities of sensor data, performing statistical analysis and extracting meaningful trends. Examples of such use include the International Continental Scientific Drilling Program (ICDP) data warehouse [ICDP 2000] which will provide access to individual data elements from selected ICDP projects.

During the past five years relational database management systems have been gradually introduced in many meteorological organizations to substitute proprietary file-based application storage concepts [Moe 1999]. Recently, a few projects are being implemented using object-relational database technology. The main drivers for the use of this technology in meteorological organizations have been the need to substitute ageing, insufficiently documented systems and the development of new web-based applications that is well supported by many modern DBMS. However, the solutions implemented are for the most based on ad hoc architecture resulting in complex and heterogeneous database landscapes. It is typical for our specific application domain, that only parts of the data are stored under the control of a DBMS while files are used for other parts. Furthermore, multidimensional modelling as applied in this project has to the best of our knowledge not been used in meteorological and climatology databases until now.

Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW'2001)
Interlaken, Switzerland, June 4, 2001
(D. Theodoratos, J. Hammer, M. Jeusfeld, M. Staudt, eds.)
http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-39/

3 Requirements and Challenges

The use of meteorological data is mainly twofold:

- Weather forecasting, where quick access to actual data is important;
- Climatology, where flexible access to high quality information about past weather is important.

As a national weather service, MeteoSwiss is responsible per legal mandate to collect and transmit to the 'Global Telecommunication System' (GTS) of the 'World Meteorological Organization' (WMO) meteorological and climatology measurements and observations from the territory of Switzerland. In order to provide meteorological services for the public (weather forecasts, alerts etc.), MeteoSwiss makes extensive use of data (observations and products) transmitted via the GTS from other countries.

The life cycle of meteorological/climatology data is shown schematically in Figure 1.

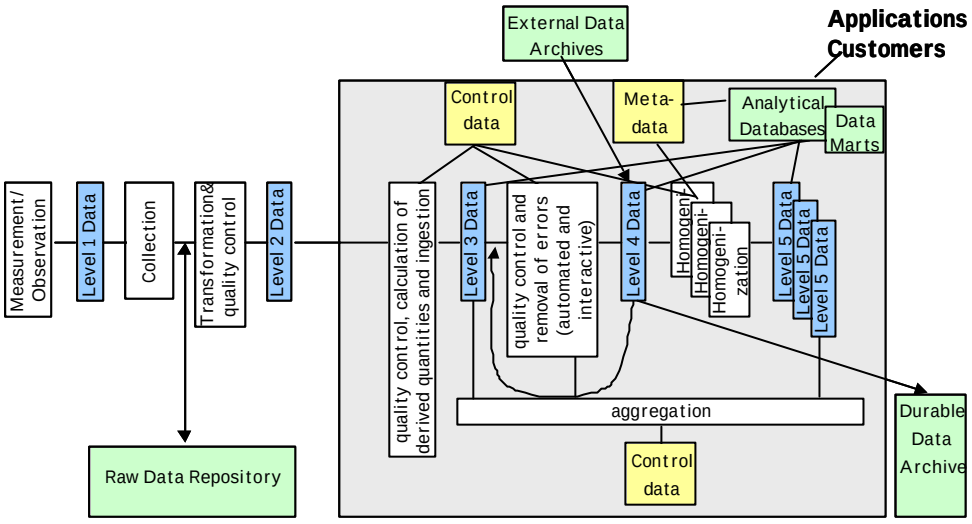


Figure 1: Schematic representation of the process chain/life cycle for processing meteorological data (adapted from [Konzelmann 1998]). The large rectangle area denotes the parts of the process that are covered by the data warehouse system.

The data undergo 5 transformation steps:

- Level 1 data are produced at the measuring site by transforming raw sensor data (e.g. voltage, counts) to physical units (e.g. temperature, precipitation rate) that are meaningful for further meteorological applications.
- Level 1 to level 2: level 1 data from various sites are collected by a polling system. Some data are subsequently transformed to level 2 data by applying calibration polynomials. Simple quality control procedures are used to flag gross errors in the data. Some derived quantities are calculated depending on the result of the quality control results. Level 1 data are stored in a 'raw data repository'. This repository will be the data source, if the transformation from level 1 to level 2 data has to be repeated using improved calibration constants. This process is called 're-evaluation' (e.g. re-evaluation of radio probe data using an improved radiation correction scheme).
- Level 2 to level 3: level 2 from different polling systems are integrated and loaded into the data warehouse system. Similar data from different polling systems are subjected to uniform quality control procedures using more rigorous tests than in the previous transformation step. Plausible values are used to calculate derived quantities (e.g. vapor pressure, mean sea level pressure). After the transformation, these data are tagged as 'level 3' data. They are further used for aggregation, propagated to analytical databases and used for customer products and services (e.g. weather forecasting).
- Level 3 to level 4: The core of this transformation step is the 'data cleansing' process [WMO 1990], [WMO 1993]. Temporal and spatial consistency checks are

added to the quality control procedures from the previous transformation steps and all checks base on an integrated dataset. Obviously erroneous data are corrected automatically and by means of interactive tools. This transformation step can be performed several times but the results of each transformation have to be stored. Cleansing of meteorological data may affect also historical data for two main reasons:

- New and more sensitive methods allow better error detection;
- Historical data (mainly digitized from publications) are added to the dataset and provide additional information to check the spatial and temporal consistency of the dataset.

- Level 4 to level 5: A selection of level 4 data is homogenized using context data (especially station histories). Homogenization in this context means to remove all non-natural fluctuations (e.g. changed calibration constants, changed instrument exposition) from the data series in order to get "a numerical series representing the variations of a climatology element with variations which are caused only by variations of weather and climate [CP 1950]". The results of this transformation step are so-called homogenization values (amounts or factors). These values produce the homogeneous data series if they are added to/multiplied with the corresponding level 4 data. This transformation step can be performed in various projects using different methods in different contexts. Each project produces an own set of homogenization values that are used depending on the customers needs.

The quality control and data cleansing procedures of all transformation steps are based on parameters stored in the 'control data'. Aggregation is performed on level 3, level 4 or level 5 data and is also based on parameters that are part of the control data. The term 'metadata' means here 'climatological context data' (e.g. information about station location, sensor types, station histories etc.).

4 The MeteoSwiss Data Warehouse System Architecture

The architecture strategy of our project is mainly a wholesale replacement of the operational systems and the establishment of a new analytical database. Operational systems in the context of our project are used to transform the data from level 2 to level 4 whereas the analytical database is the classical 'data warehouse'. In the whole design much emphasis is placed on the design of the operational systems.

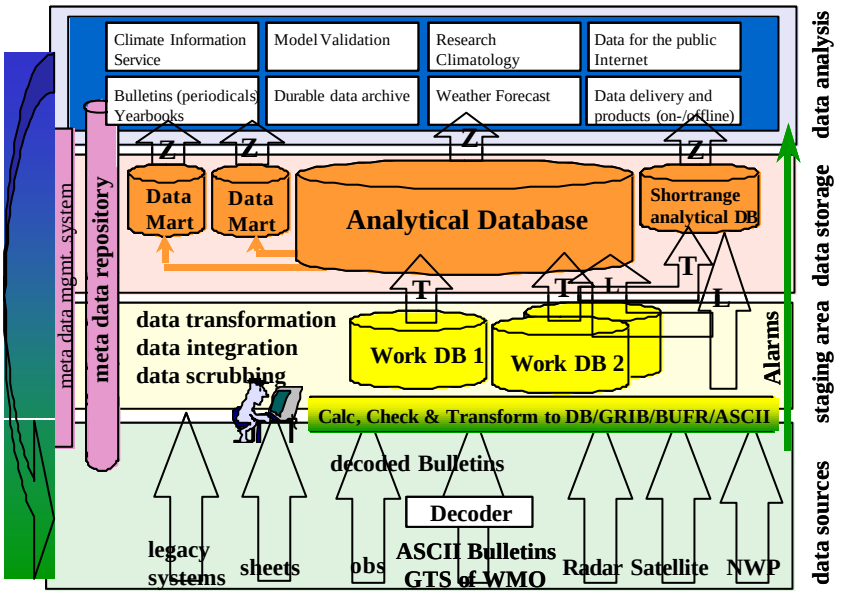


Figure 2: Conceptual architecture of the MeteoSwiss data warehouse system. Arrows depict data flows (T - transformation and loading, L - loading without transformation, Z - ad hoc queries).

The conceptual architecture proposed in our project consists of four layers (see figure 2). This architecture is not a fully conventional data warehouse. It is rather a special case where DW technology is combined with classical relational database technology. This approach is more flexible to cope with special cases. The characteristics of the four layers are described in the next sections.

4.1 Operational Data Sources

There is a wide variety of sources for meteorological and climatology data:

- Observation systems operated by MeteoSwiss and partners usually provide data in high temporal (10' or less) and spatial resolution (e.g. 115 automatic weather stations for Switzerland) for a very limited spatial area;
- The numerical weather prediction model of MeteoSwiss produces results in hourly resolution for 51 parameters on 385x325 grid points at 45 vertical layers;
- The GTS of WMO and similar systems provide data from all over the world in a usually low temporal resolution (hourly or coarser).
- Various data archives (databases, file systems etc.) provide historical data that are very important for climatology research (e.g. climate change research).

The data from the sources have different types:

- Point measurements from automatic weather stations and weather stations providing hand-written protocols;
- Profile measurements from upper-air stations and measurements along trajectories;
- Picture data and field data from satellites and weather radars;
- Volume data from weather radars and numerical weather forecast models.

4.2 Staging Area

The staging area comprises all databases and tools that are necessary for digitizing, integrating, loading, quality controlling and cleansing the meteorological data. Since cleansing meteorological data may also affect historical data it is necessary to store comprehensive datasets in the so-called work DBs. Special emphasis is given to the development of data quality control systems and tools for error correction. The work DBs contain data at lowest level of granularity and some lightly summarized data.

4.3 Storage Area

This is the classical 'data warehouse'. In the analytical database, the meteorologists and climatologists of MeteoSwiss will find an integrated set of data to support their decisions in weather forecasting and climate research.

Since data quality is less important compared to quick access for certain applications, a 'short range analytical database' will be implemented which will be fed directly from the data source layer.

4.4 Analysis Area

This layer contains various analysis and data visualization tools, which are not in the scope of the present project. Various interfaces are provided to access the storage area.

5 Modeling Climatology Data

The modeling process used in our project is based on the classical database modeling process distinguishing between conceptual, logical and physical models extended by the implementation model approach proposed by [Martyn 2000]. According to this approach, the logical model is an ideal relational model that is the input to an implementation model where performance enhancement aspects such as vertical table partitioning are considered. An important consideration for using this approach is to support transparency in logical design e.g. by explicitly designating a partitioning step in the design process. The relevant characteristics of data warehouses include storage of materialized views and de-normalized, redundant data tables. These performance enhancement schema elements are best defined as part of the implementation design. An added benefit of this approach for our case is that expert end-user involvement (which is necessary due to scientific know how needed for the design of the data model) can be extended into logical data modeling and the logical data model can be used as a vendor-independent development specification allowing DBMS-specific optimization.

The conceptual model covers all aspects of the data without considering the technical database realization. Specifically the conceptual data model has been used as a map for defining boundaries for subsequent implementation projects. It has proven a powerful communication instrument within the project team and for external project communication.

The conceptual model of the system has been developed through a series of coached workshops with extensive user involvement and through reverse engineering of existing databases followed by reviews of the results. Important challenges during the conceptual modeling where:

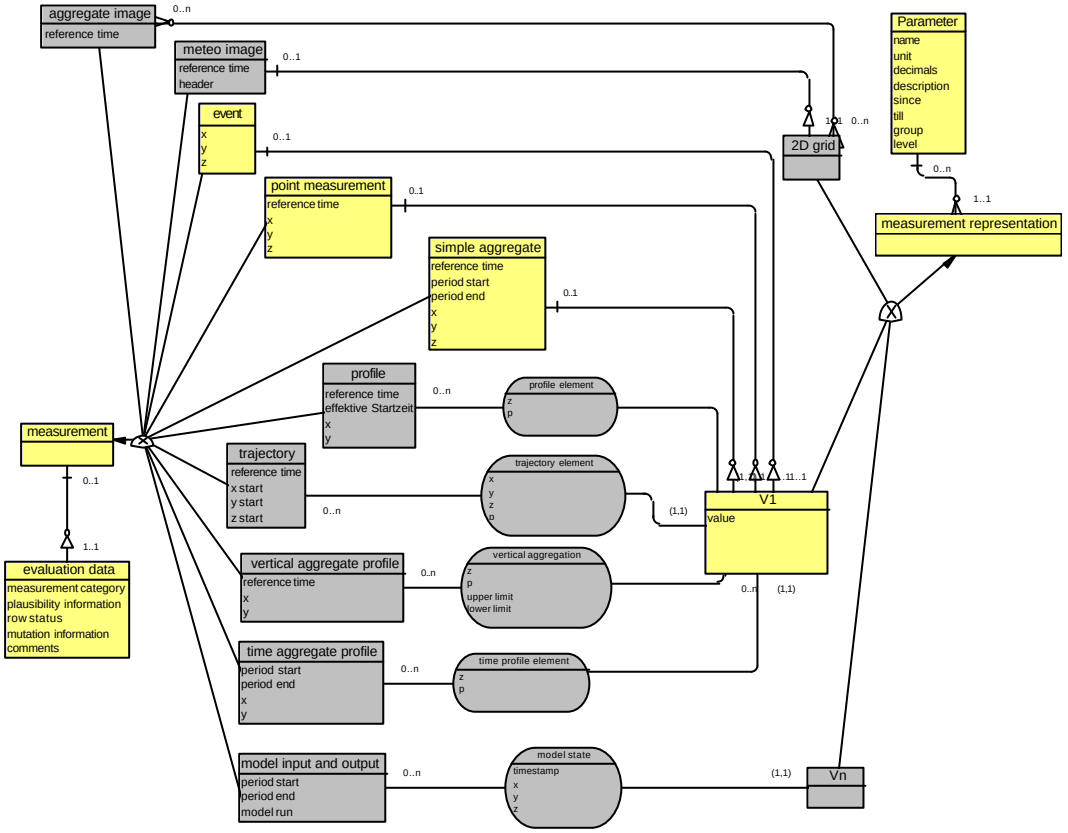
Establishing a common vocabulary of the application domain: existing database systems were inadequately documented and no consistent terminology and naming was enforced. An important result of the conceptual modeling has been the consolidation and precise definition of used terms.

Abstracting from current implementations and legacy systems: often and due to lack of documentation it proved difficult to extract conceptually relevant information. The resulting conceptual model has provided a complete and relatively simple overview of a very complex database landscape. As an example, the following figure depicts the conceptual representation of more than 1000 different kinds of meteorological measurements currently performed.

The conceptual model distinguishes between two main groups of entity types: measurement data for the various meteorological phenomena and climatology context data (metadata) required for their correct interpretation.

Measurement data are represented in the subtypes of entity type “Measurement” in Figure 3.

Figure 3: Example of conceptual model of meteorological measurements (tool-specific notation). The model represents more than 1000 different measurement types currently performed. These include point data, time series



data, trajectory data and 3-dimensional multi-vector numerical weather model data. Measured data pertain to meteorological “parameters” which correspond to measured real-world phenomena.

It is important to note, that measurement data consists in general of a value, a representation and data regarding the type of measurement made. Different values may be in fact aggregated over space and/or time resulting in different types of measurements. Measurements can be interpreted only by considering their context data.

Context data can be divided into description of the physical measurement environment and information required for operations on measurement data (e.g. homogenization, quality control). We note at this point that the boundaries between the different kinds of context information are not hard.

The important role of the context information lies in the fact that it defines the classification dimensions for measurement data. However, certain complexities had to be considered:

The conceptual model is not the basis for the implementation of a specific data warehouse system. Rather, it describes data that are managed and used by various applications and in various databases. As such a multidimensional modeling technique optimized for OLAP (e.g. [SBHD 1998] was not explicitly used; candidate classification dimensions where informally documented in the model.

Large parts of the conceptual model for context information do not fit a multidimensional model. Control information for aggregation, homogenization and test functions are among these parts.

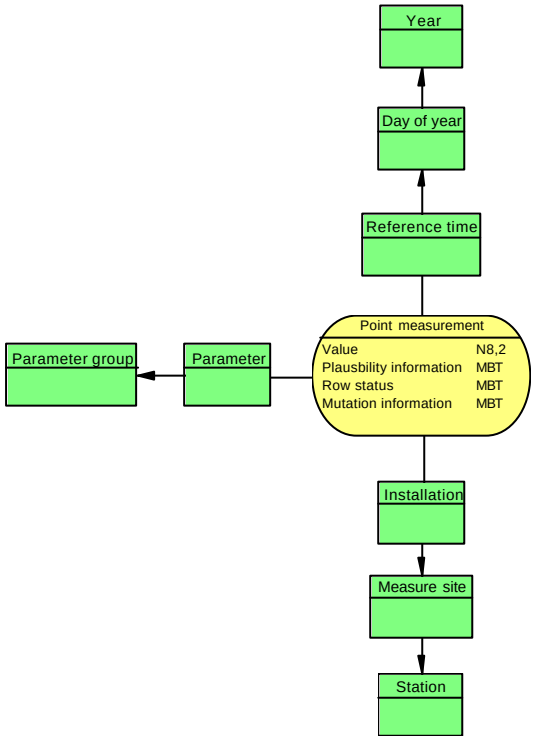


Figure 4: Example of ME/R Model for point measurement facts. The dimensions are measured parameter, installation where measurement took place and reference time of measurement.

Due to this fact, multidimensional modeling has been introduced in the subsequent modeling process for selected implementation projects. For these, combined E/R [Chen 1976] and M/ER Model [SBHD 1998] concepts are used. An example of ME/R modeling is presented in Figure 4. It depicts the modeling of point measurement data and uses an adapted notation but without modifying the semantics of the metamodel of M/ER. More specifically facts are depicted by an n-ary relationship (rounded rectangle) and classification dimensions are modeled as entity types. Possible classification relationships are depicted by arrows; note however, that specialized aggregation operators should be applied along each classification dimension and between each classification step. Snowflake schemas can implement the above conceptual model. The large numbers of facts in the database (ca. 1000 parameters, 10-minute time intervals for ca. 100 years and several hundreds of installations) require performance enhancement techniques including table partitioning.

6 Conclusions

In the described MeteoSwiss project we have developed an architecture that, while not a traditional data warehouse, uses several elements from data warehouse technology. We support the entire data life cycle and enforce a unidirectional data flow which allows us to optimize each data storage for the use patterns it should support (OLAP vs. OLTP). This approach promises to solve the long-term performance issues which often plague meteorological and climatology databases.

References

[Chen 1976] P. Chen, The Entity Relationship Model - Towards a Unified View of Data. ACM TODS, 1:1, 1976.

[CP 1950] Conrad, V. and L.W. Pollack, 1950: 'Methods in Climatology', Harvard University Press.

[ICDP 2000] ICDP Clearing House, http://icdp.gfz-potsdam.de

[Konzelmann 1998] Konzelmann, Thomas, Martin Kiene, Rudolf Doessegger and Gabriela Seiz, 1198: NEW TREATMENT OF REAL TIME CLIMATE DATA SETS FROM SMI WEATHER STATIONS. Proceedings 2nd European conference on Applied Climatology 19-23 Oct 1998.

[Martyn 2000] T. Martyn, Implementation Design for Databases: The "Forgotten" Step, IEEE IT Professional, Macrh/April 2000, pp. 42-49.

[Moe 1999] M. Moe (Ed), Proceedings from the Oslo EUMETNET - ECSN Workshop, 11.-12. October 1999.

[SBHD 1998] C. Sapiam, M. Blaschka, G. Höfling, B. Dinter, Extending the E/R Model for the Multimensional Paradigm, Proc. International Workshop on Data Warehouse and Data Mining (DWDM, in connection with ER), November 1998.

[WMO 1990] WMO, 1990: Guide to climatological practices. 2nd edition. WMO 100.

[WMO 1993] WMO, 1993: Guide on the Global Data-Processing System. WMO 305.