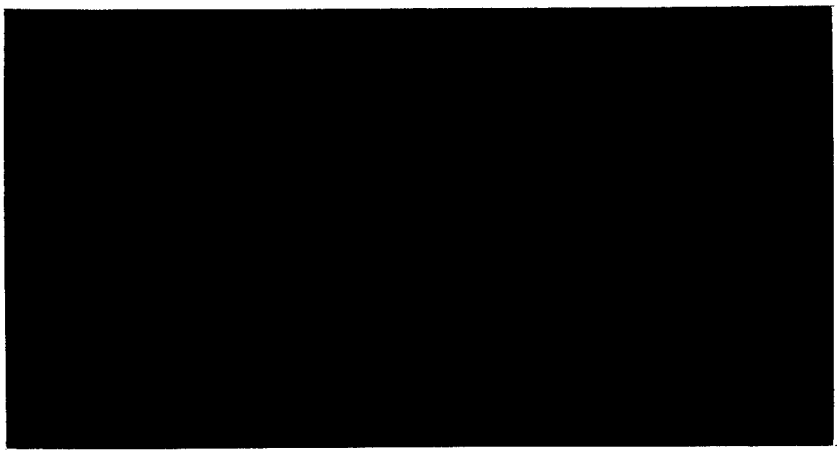

DEPARTMENT OF THE AIR FORCE

AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT/GCS/ENG/00M-09

DATA WAREHOUSE TECHNIQUES TO SUPPORT GLOBAL ON-DEMAND WEATHER FORECAST METRICS

THESIS

Meriellen C. Joga, Captain, USAF

AFIT/GCS/ENG/00M-09

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense or the U. S. Government.

AFIT/GCS/ENG/00M-09

DATA WAREHOUSE TECHNIQUES TO SUPPORT GLOBAL ON-DEMAND WEATHER FORECAST METRICS

THESIS

Presented to the Faculty

Department of Electrical and Computer Engineering

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Computer Systems

Meriellen C. Joga, B.S.

Captain, USAF

March 2000

AFIT/GCS/ENG/00M-09

DATA WAREHOUSE TECHNIQUES TO SUPPORT GLOBAL ON-DEMAND WEATHER FORECAST METRICS

Meriellen C. Joga. B.S.

Captain, USAF

Approved:

Henry B. Potoczny

Henry B. Potoczny, Ph.D.
Committee Chairman

8 MARCH 2000.
date

Ronald L. Lowther

Ronald Lowther, Ph.D., Lt Col
Committee Member

23 FEB 00
date

Cecilia A. Miner

Cecilia A. Miner, Ph.D., Lt Col
Committee Member

8 March 2000
date

Michael L. Talbert

Michael L. Talbert, Ph.D., Major
Committee Member

8 Mar 2000
date

Acknowledgments

Over the past 18 months, I've depended upon the support and assistance of many friends and colleagues. I am especially grateful to my fellow students and their spouses. Knowing that I had a support network that was willing to provide any help I required gave me the peace of mind to focus on other priorities.

I'd like to thank Major Talbert for all of the reference material he loaned to me and for being such an inspiration. His high expectations and encouragement provided the motivation I needed to persevere, especially towards the end of my research. Dr. Potoczny, you are by far the best lecturer I've ever had the pleasure of observing. Your classes were always a welcome respite from the drudgery of AFIT. Thanks also go to my partner-in-crime, Capt Darryl Leon, and to my 'weather' advisors, Lt Col Minor and Lt Col Lowther, for their insight into the specifics of terminal aerodrome forecast metrics and the weather domain in general.

Most importantly, I would like to express my appreciation to my wonderful husband, Al, and my son, Alek, for their continuous love, encouragement, and understanding despite the separation, stress, and other difficulties that we've been through over the last year and a half. I would also like to thank my mother for listening sympathetically to all of my worries, concerns, and other musings whenever I needed a friend and confidant.

Meriellen C. Joga

Table of Contents

List of Figures

Page

Abstract

Air Force pilots and other operators make crucial mission planning decisions based on weather forecasts; therefore, the ability to forecast the weather accurately is a critical issue to Air Force Weather (AFW) and its customers. The goal of this research is to provide Air Force Weather with a methodology to automate statistical data analysis for the purpose of providing on-demand metrics. A data warehousing methodology is developed and applied to the weather metrics problem in order to present an option that will facilitate on-demand metrics. On-line analytical processing (OLAP) and data mining solutions are also discussed.

DATA WAREHOUSE TECHNIQUES TO SUPPORT GLOBAL ON-DEMAND WEATHER FORECAST METRICS

I. Introduction

Air Force pilots and other operators make crucial mission planning decisions based on weather forecasts; therefore, the ability to forecast the weather accurately is a critical issue to Air Force Weather (AFW) and its customers. Erroneous forecasts can waste precious manpower, time and money, but they also have the potential to seriously affect flight safety, placing the aircrews' lives in jeopardy. Unfortunately, Air Force Weather has seen a steady decline in mission effectiveness due to the lack of experienced forecasters as well as the increased tempo of operations around the globe as documented in the executive summary of the Systems Requirements Document for the Reengineered Air Force Weather Weapon System (2:Executive Summary). As a result, AFW has developed and implemented the Air Force Weather Strategic Plan which is aimed at reengineering their business processes.

1.1 Importance of Research

One of the primary improvement areas that the Air Force Weather Strategic Plan addresses is the implementation of an operator-focused Air Force Weather metrics program to monitor the operational, technical, personnel, and resource health of the Air Force Weather system (23:29). A consistent method of evaluating the skill level of AFW forecasters is required to plan for and assess specific improvements in Air Force Weather's business processes. Currently, forecasting data is collected and formatted manually, then

typed into a computer as a pre-formatted string and forwarded through a message service to a centralized location. These messages are then scanned into a database and erroneous data is discarded. This system doesn't interface with any records of actual observations; therefore, any metrics tracking the reliability or quality of the forecasts have to be generated manually at forecasting locations. Although some of this process used to be automated, that system has since been cancelled for financial reasons. The only metric currently collected is based on the visual flight rules (VFR) to instrument flight rules (IFR) threshold of 1,500-foot ceilings and 3 miles of visibility. It would take approximately 6 months to initiate data collection for a new metric under the current system.

1.2 Problem Statement

Two years ago Brigadier General F. P. Lewis, Air Force Director of Weather, shared his vision of on-demand weather forecasting metrics (10; 23:Operations Focus). The delay in the then current manual process being used was unacceptable in view of the advanced technology available. What is required to solve the problem is a method of verifying terminal aerodrome forecasts (TAFs) that is flexible and adaptable to change, yet provides a timely response to newly developed metrics.

1.3 Current Initiatives

The new Air Force Weather Weapon System (AFWWS) (2) is one of the initiatives that will completely revamp the automation of weather data collection, eliminating the current stovepipe systems such as the Automated Weather Distribution System (AWDS) and incorporating a centralized forecasting process known as a weather *hub*. However,

although a metrics program is a major action area, it is only addressed in very general terms in the system requirement document. "The AFWA PS [Air Force Weather Agency Production System] shall automatically compile the data required to execute the AFW metrics program" (2:Sec 3.2.3 p 78). The system requirements document goes on to state "AFW metrics are defined in AFI 15-114." Unfortunately, AFI 15-114 is prescriptive in nature and doesn't provide a description of the intended metrics program. It simply states that each unit is to develop metrics that are specific to their individual requirements (6:Sec 3). In this light, a very general approach to metrics collection will be required.

The goal of this research is to provide Air Force Weather with a methodology to automate data analysis for the purpose of providing on-demand metrics. This is accomplished through review of the data format and statistical analysis requirements. Data warehousing, on-line analytical processing (OLAP), and data mining solutions are all examined in order to present an option that will facilitate on-demand metrics.

1.4 Document Organization

The next chapter provides background information used in this research effort. It introduces data warehousing, OLAP, and data mining concepts. Data marts are also discussed as an alternative to full-scale data warehousing. Chapter 3 details the steps taken in designing an appropriate solution. Chapter 4 contains a description and analysis of the data mart solution considered. The last chapter summarizes the conclusions that were reached, as well as recommendations for future work in this area.

II. Background Material

On-demand dynamic metrics production requires immediate access to the underlying data and a means to analyze it. Previous efforts utilized the Terminal Aerodrome Forecast Verification (TAFVER) programs written specifically for the purpose of calculating metrics. Unfortunately, the underlying data stores were incomplete and weren't readily accessible by the end-user due to the limitations of the available technology at the time these systems were developed. Data warehousing is a relatively new solution to the data access problem that could be exploited in order to produce on-demand, dynamic metrics. A derivative of data warehousing called data marts is an alternative form of the technology that requires much less effort in terms of time and money. The capabilities and advantages of data marts are also worth exploring in terms of metrics production.

2.1 Terminal Aerodrome Forecast Verification (TAFVER)

TAFVER was an automated program designed to measure the quality of weather forecasting support provided by the weather community. Variations of the TAFVER system were used from the early 1970's until the funding for the program was terminated in 1998. The first version of TAFVER was created in the early 1970's and provided the capability to verify ceiling and visibility forecasts based on hourly observations (13:1-28). TAFVER II, created in the early 1990's, added additional TAF elements, verification of combinations of TAF elements, user-specified category thresholds, and a wide variety of grouping methods (7:iv). A few years later, it became necessary to convert TAFVER II from DB2 to ORACLE. The decision was made to construct a new system in conjunction

with the conversion (25:5). However, funding for the TAFVER program was eliminated before this version was implemented.

2.2 *Advanced Weather Interactive Processing System (AWIPS)*

The National Weather Service uses AWIPS to analyze, process, and display hydrometeorological data and to disseminate warnings and forecasts in a rapid, highly reliable manner (1). Recent enhancements to AWIPS include the capability to automatically save data to be sent to the National Centers for Environmental Prediction (NCEP) for comparison of forecasts to observations to determine the accuracy of the forecasts based on some standard. Verification metrics are calculated on 8 weather elements from approximately 95 stations in the contiguous United States. The elements verified for aviation forecasts are ceiling height, visibility, and wind speed and direction.

2.3 *Data Warehousing*

Data warehousing is one of the more frequently extolled developments in the computer industry today. Although each subject matter expert defines data warehousing uniquely, Ralph Kimball's[1] is relatively consistent with most of the others. According to Mr. Kimball, a data warehouse is "a copy of transaction data specifically structured for query and analysis" (15:310). Simply stated, it's a means of storing historical data so that information can be efficiently calculated and retrieved. A data warehouse is a physically

[1] "Ralph Kimball was founder and CEO of Red Brick Systems. He is a leading proponent of the dimensional approach to designing large data warehouses. He currently teaches data warehousing design skills to IT groups and helps selected clients with specific data warehouse designs." (16)

separate store of data transformed from the application data found in the operational environment (see Figure 1 below). It supports information processing by providing a reliable source of historical data from which to do analysis.

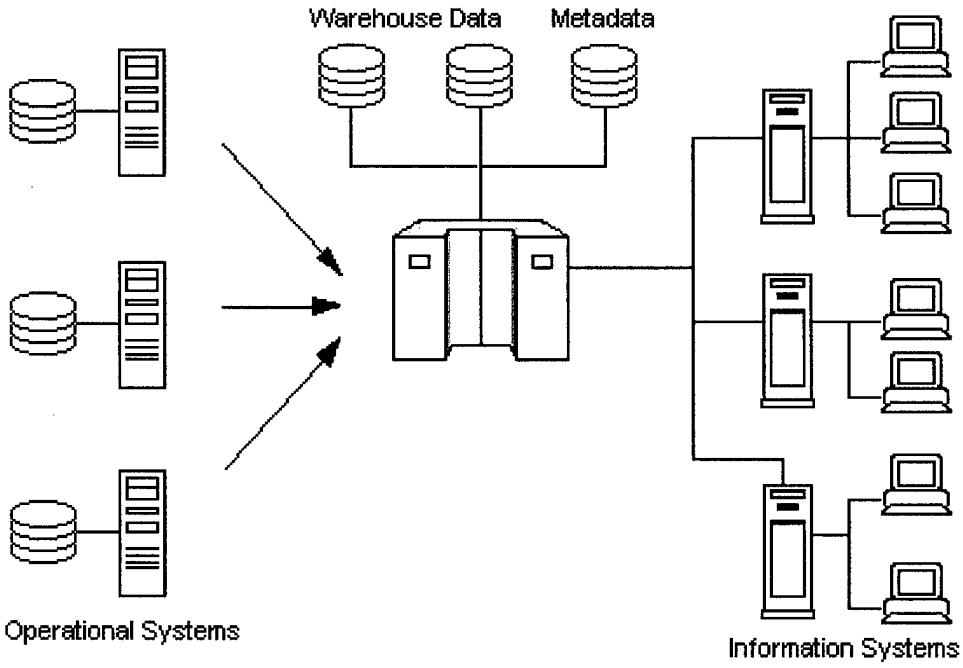

Figure 1 - Data Warehouse (26:3)

2.3.1 Characteristics

In his book, "Building the Data Warehouse," W. H. Inmon[2] lists the primary characteristics of a data warehouse as subject oriented, integrated, non-volatile, and time variant (12:33-43). These characteristics distinguish the data warehouse from a typical database. Traditional databases are typically functionally oriented, supporting business activities. A data warehouse, on the other hand, focuses on a particular subject such customers or sales. Inmon argues that the most important characteristic of a data warehouse is its integration. Multiple sources of data may have conflicting formats and rules for the same data item. For example, one data source may store a specific boolean value as 'yes' or 'no'. (One example of a boolean value is whether or not a line of a forecast is a temporary weather condition.) Another data source would store the same information as 'true' or 'false' while a third could use '0's and '1's. The data warehouse must integrate the data from the various sources in such a way that these inconsistencies are resolved. The third characteristic is that it is nonvolatile. Instead of storing the most current information, a data warehouse would contain records of all of the transactions that affect this information so that the *current information* could be reconstructed for date in the past. Once data is entered into the data warehouse, it seldom changes. This leads us to our final property. The element of time is unique to data warehousing. An operational data store only reflects the most current state of the data. Data in a data warehouse is time

[2] "W.H. Inmon is a consultant and vice President of prism Solutions, a company that automates the development and operation of data warehouses. He has 31 books in print, speaks frequently at international conferences, and is a regular contributor to professional journals." (12: back cover)

variant in the sense that the state of the business over any given period of time can be reconstructed from the data stored in the warehouse.

2.3.2 Benefits

Some of the more pervasive phrases used to describe benefits that can be attained through data warehousing embrace topics such as converting data into business intelligence, making management decisions based on facts instead of intuition, getting to know the customers better, and gaining competitive advantage. These are all very wonderful, lofty goals that may be facilitated by data warehousing but don't reflect the reality of what data warehousing can achieve independently. Rather, data warehousing is a valid data storage method that can generate a large return on investment if it's implemented effectively in conjunction with efficient data retrieval methods. Benefits that can actually be attained are as follows:

1. The burden of data analysis and information extraction is removed from the operational system.
2. Query and reporting capabilities are optimized.
3. A user-friendly environment for query and reporting which minimizes the knowledge and time required for generating new queries/reports is provided.
4. A repository of "clean" (correct) data that doesn't require fixing the operational system is provided.
5. Access to data from multiple systems and to external data is enabled.
6. A repository of historical data is provided.

7. Security is enhanced. Personnel who only need query access are prevented from tampering with the transaction data and logic.

2.3.3 Drawbacks

Any list of benefits also has associated drawbacks or shortcomings. Data warehousing is no exception. Development of a data warehouse takes a long time, and according to Douglas Hackney[3], can be very expensive ($2 to 5 million) (11:4). In order to be successful, cooperation and communication across the entire business is required along with high levels of sponsorship, long term political will, determination, faith, etc. Last but not least, development personnel must have an established set of skills which are in short supply in today's computer industry.

2.4 Data Marts

A data mart is a scaled down version of a data warehouse designed with a specific purpose in mind. It may be either a distinct entity or part of a larger data warehouse (see Figure 2). Because a data mart is smaller and has a very distinct focus, it can be developed faster and requires less space than a full-blown data warehouse. Therefore, the cost is reduced significantly. Because the data mart is smaller, with fewer users, one time costs for disk storage, network communication, and software tools are decreased. The simplicity of a data mart design also cuts down on recurring expenses such as the integration and

[3] "Douglas Hackney is president of Enterprise group Ltd., a consulting and knowledge transfer company specializing in data warehousing and data marts. Mr. Hackney has over seventeen years of experience in business management and in designing and implementing information delivery system solutions for Fortune 500 organizations. He is a frequent speaker at data warehouse industry conferences, is a master instructor and the primary instructor for data mart initiatives for The Data Warehouse Institute, and is a founding board member of The International Data Warehouse Association." (11: back cover)

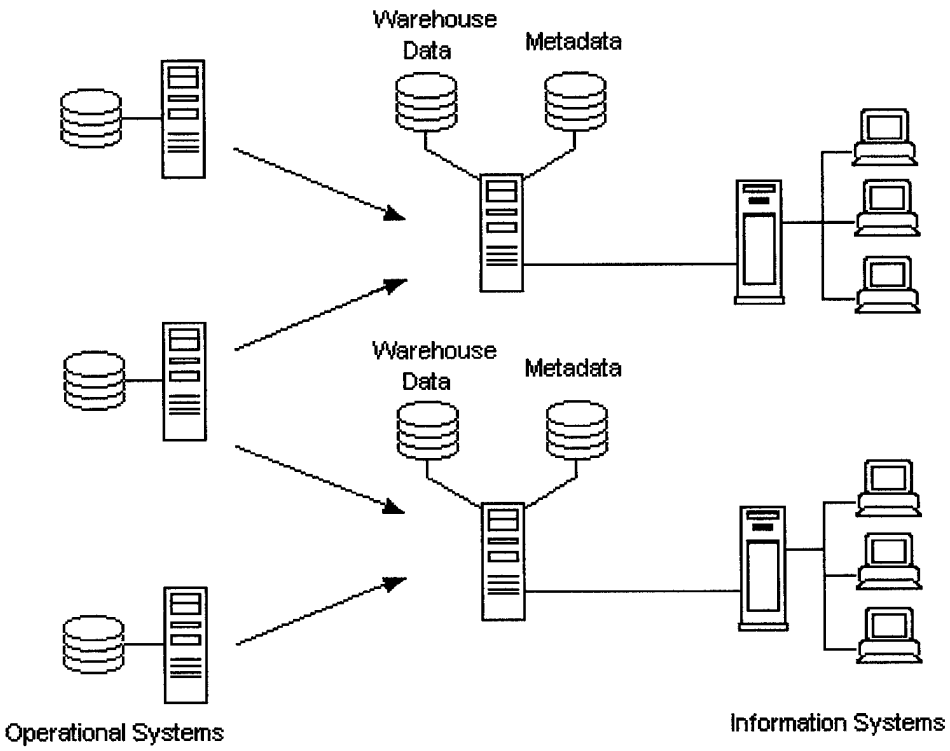

Figure 2 - Data Marts (26:8)

transformation of data and metadata maintenance. Since the single purpose of a data mart limits the need for communication and cooperation between departments, much of the technical and political risk inherent in the development of a data warehouse is reduced or eliminated. Limiting the scope of a data mart also allows lower levels of sponsorship. Because fewer departments are involved, decisions can be made at lower levels of management. Since data marts are smaller and limited in scope, they also present a good environment for development of necessary skills. In short, data marts have fewer complexities than a data warehouse and require less time, money, management involvement, and established skills.

2.5 *On-line Analytical Processing (OLAP)*

The real benefits of a data warehouse are not achieved through use of the data warehouse alone but, more significantly, from the means by which we extract the information it contains. OLAP gives data analysts the capability to transform large amounts of data into meaningful information. In the past, management has often made decisions based on *gut feel* and experience rather than solid analyses and tested hypotheses. Analysts can utilize OLAP tools to validate or disprove their theories, which in turn enables management to make better-informed decisions quickly.

OLAP provides a multidimensional conceptual view of the data that makes queries of large amounts of data much more efficient. Traditional methods of database access only affect a relatively small number of records per transaction. OLAP, on the other hand, typically searches the majority of the records in the applicable tables in order to isolate and calculate the requested information. For this reason, traditional query techniques are extremely inefficient at accessing the information stored in a data warehouse. Another limiting factor concerns the manner in which the data must be processed in order to transform it into meaningful information. Traditional data retrieval engines simply don't have the functionality that's required to transform the data efficiently.

One of the most significant capabilities of OLAP is the ability to drill down into the information and conversely, drill up to a higher level of abstraction. An analyst may initially ask for information based on specific groupings such as by organization and by month. The results may provoke questions about the underlying detail. With drill-down, the view can easily be changed to data for a specific organization or to be grouped by day.

Drill-up shows a higher level view such as summing the data by year instead of by month. These capabilities provide OLAP much of its flexibility and power.

2.6 Data Mining

According to Avi Silberschatz and Stan Zdonik, data mining is one of the many research areas that is "breaking out of the box" (24:49). Data mining is the process of automatically extracting valid, useful, previously unknown, and ultimately understandable information from large databases and using it to make business decisions. It promises an easy to use, understandable approach, with the software making the choices and calculations for the user; however, the process of data mining requires substantial human effort and interaction. Off-the-shelf applications entice with promises of ease of use, but in most cases computer systems expertise and a thorough knowledge of the data are necessary for productive analysis. Data mining has the power to deliver what it promises but the effort it requires isn't necessarily worth the investment.

III. Methodology

The methodology presented in this chapter represents a synthesis of the varying methodologies offered by recognized data warehousing experts. The available literature was examined to identify the *best-of-class* methods as they applied to the weather forecast metrics problem. These methods were then integrated to form the methodology presented below. The results described in chapter IV demonstrate the successful application of this integrated methodology.

Although data warehousing is becoming a popular solution, because it is a relatively new and broadly extensible concept, the details of the modeling techniques aren't yet firmly defined. Unlike software engineering, standard lifecycle models for data warehouses have yet to be developed. Among the traditional lifecycle models, the spiral model of development is the best fit. However, the steps in developing a data warehouse are intermingled and overlapping versus sequential as the spiral model implies. In the opinion of this researcher, Hackney's three dimensional business-driven spiral model is a better representation for the data warehouse lifecycle. This adaptation of the traditional spiral model focuses on business-driven requirements and balancing the various issues and challenges inherent in data warehouse development. Figure 3 depicts Hackneys model.

Subject matter experts have yet to establish a common methodology to be used during data warehouse development. Their methodologies, like their data warehouse definitions, are all unique in some fashion. Each subject matter expert seems to have created his or her own development process. Their work tends to focus only on the specific

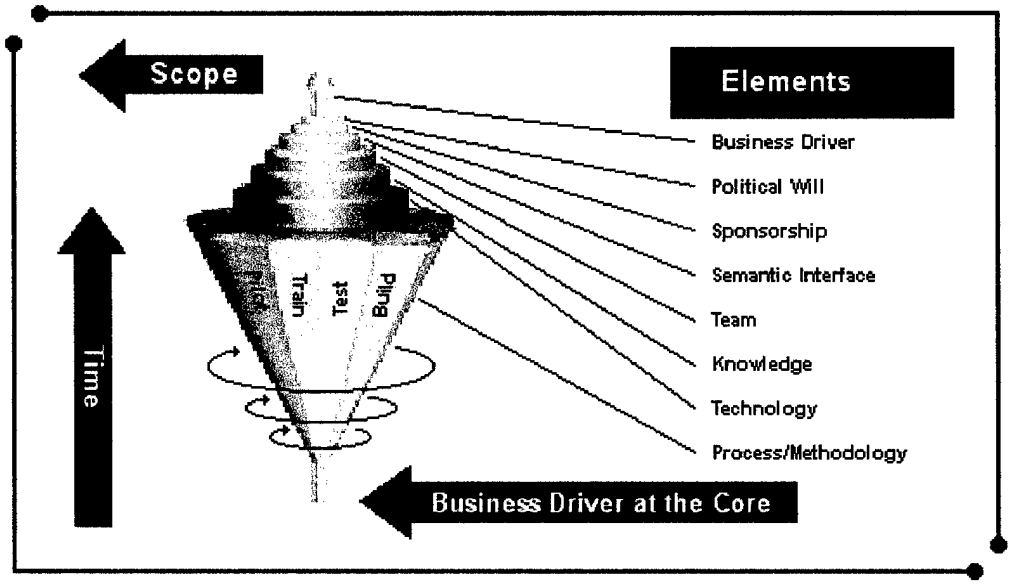

Figure 3 - Business Driven Spiral Model (11:148)

aspects that they find important or confusing, describing the remainder of the process only superficially, if at all.

This chapter describes the synthesis of those multiple methodologies as it was applied to the Air Force Weather metrics problem described in chapter 1. The detailed knowledge that contributes to a successful data warehouse could fill several books. Therefore, this chapter presents an overview of the process, with specific AFW requirements in mind. The following key steps are reviewed along with several significant issues: problem definition, requirements, modeling, design, data extraction, data analysis, and purging extraneous data. It is important to note that these steps should be performed concurrently when possible and that the entire process is iterative. Even after the initial project is complete, a successful data warehouse will continue to develop and grow to meet the organization's changing needs.

Accompanying any data warehouse development effort is a series of decisions that have to be made and issues that need to be addressed that don't fit neatly into any one specific step of the process. Some of the issues such as political will and sponsorship don't necessarily relate to any of the steps in the process but nonetheless, must be resolved. AFW is not exempted from these decisions even though their organizational structure and financial planning process are considerably different than those of a conventional commercial enterprise. These issues are discussed near the end of the chapter.

Note that the terms *data warehouse* and *data warehousing* have been used exclusively in this chapter in lieu of *data mart* and its associated terminology. The methodology described herein can be applied to either a data warehouse or its descendant, a data mart.

3.1 Define the Problem

The first step in any troubleshooting effort is to define the problem. This is especially important in an information systems environment. Often, in their enthusiasm for the latest technology, information systems team members attempt to apply new technology, such as data warehousing, without targeting a true business-related need. This leads directly to solutions looking for problems, which wastes manpower, time, and financial resources.

It is worth the time to put the problem definition in writing even if it appears to be intuitive or self-evident. Not only will it help to narrow the scope of the problem, but also later reference to the problem definition can help keep the focus on the identified business

needs. This will be critical as development of the data warehouse progresses because it gives a specific focus to the project. If that focus is adhered to, requirements creep along with all of its associated issues such as missed deadlines and increasing budgets can be minimized.

3.2 Gather Requirements

Once the scope of the project has been narrowed through the problem definition, the development team must survey user requirements. As with many of the steps in data warehouse development, requirements gathering can be difficult and elusive. Users aren't always sure of what they want until they've seen what technology can offer. Douglas Hackney suggests asking the users a standard set of questions relating to specific areas of concern (11: 165). The six core questions he recommends are:

1. Subject area (What do you need to know about?)
2. Atomic level of fact detail (What level of detail do you need?)
3. Length of fact detail history (How far back in time do you need that detail?)
4. Required business dimensions (How do you like to view, or 'slice,' the business? By product? By customer?)
5. Multidimensional aggregation requirements (What combination of 'views,' or 'slices,' is valuable to you in a report or analysis? Sales by customer? By product?)
6. History tables (What relationships do you need to capture, track, and/or trend?

The answers to these questions should lead to a workable set of requirements from which a basic data warehouse can be developed.

Unlike traditional software development, the requirements for a data warehouse may never be fully defined. According to Arbor Software, "Users are rarely satisfied, and once they get an initial application, they are almost certain to want more information, greater data accuracy, new calculations, additional ways of presenting the data and faster performance" (22:section 6). It's important to distinguish between newly discovered requirements that should be incorporated into the current development effort and those that should be postponed until the next iteration.

3.3 Modeling Data

A data model is a graphical representation of the organization of the data along with detailed descriptions. It helps the development team conceptualize the data so they can design the warehouse. It's also a means of identifying and resolving various problems, such as reconciling dissimilar units of measurement and unrelated encoding practices, which occur when combining data from disparate sources. Areas to take into consideration while modeling the data are as follows:

3.3.1 Enterprise Data Model

If one exists, the enterprise data model can be an invaluable asset as it encompasses all of the data available to the business. It defines a consistent view of each of the data elements the business uses along with the sources of that data. Many of the issues the data

warehouse development team must contend with should have been either resolved or at least identified in the enterprise data model. Difficulties encountered during data integration are consequently easily solved as a result of the common business rules and semantics established by the enterprise data model. For a large data warehouse effort, the time spent up front developing a full-blown enterprise data model can prevent numerous data integration errors in the data warehouse model.

3.3.2 Relational Versus Multidimensional Versus Hybrid Models

The types of data models normally considered for data warehousing include relational, multidimensional, and hybrid models. The debate arguing the benefits and disadvantages of the different types of models as they apply to data warehousing is by no means resolved. Traditional methods such as relational models have the advantage of being familiar, proven technology that follow widely recognized standards. A relational database provides support for large amounts of data; however, the functionality and performance of the associated data retrieval and analysis tools are minimal when compared to those developed for the multidimensional models.

Multidimensional models are considerably more efficient at analyzing data because much of the information is pre-processed and stored in the database. Because of this pre-processing, however, the volume of data is enormous and the processing required to populate the hypercube can be burdensome. Figure 4 depicts a multidimensional hypercube. This type of model is best utilized for small amounts of data that must be accessed repeatedly and efficiently. One practical use is to maintain the previous year's statistics separate from the larger data warehouse for immediate referral. The details of

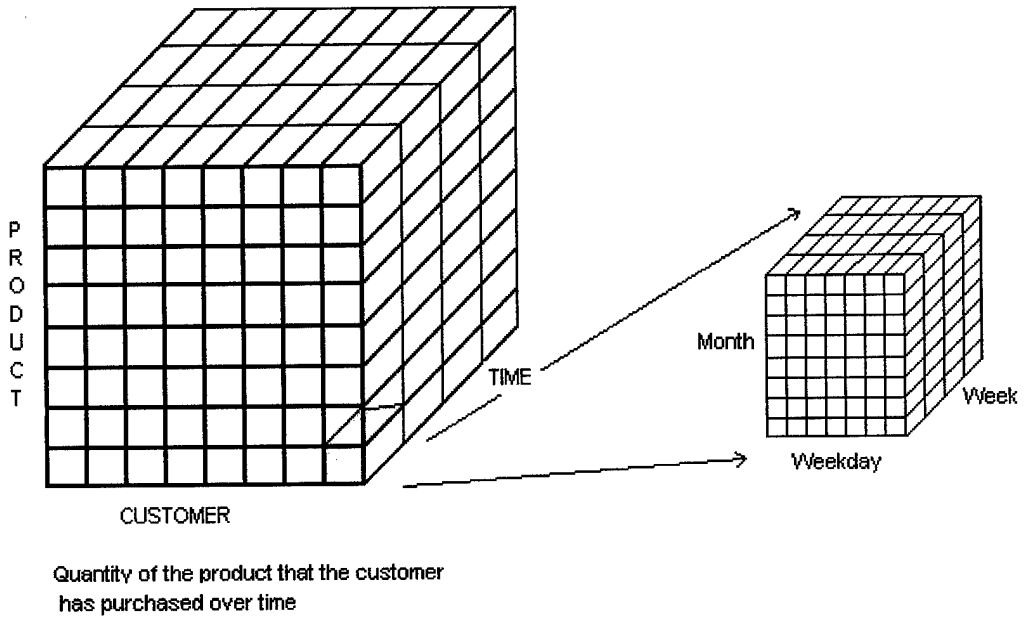

Figure 4 - Multidimensional Hypercube

how this type of model is instantiated are typically proprietary and very dependent upon the particular vendor providing the solution.

Hybrid models, such as star schemas, are a promising development that is popular among the data warehousing experts. A star schema differs from a relational model in that it is denormalized. A normalized database stores a specific piece of information such as a customer's address in only one location. A denormalized database may store that information in several different locations in order to speed up access times. The foundation of a star schema is its primary fact and dimension tables. The fact tables contain the low-level detail of interest to the user. The dimension tables encompass the various ways to group and compare that data and details about those groupings. The star schema may also include additional tables that contain pre-processed information much like a multidimensional model. Figure 5 depicts a basic star schema.

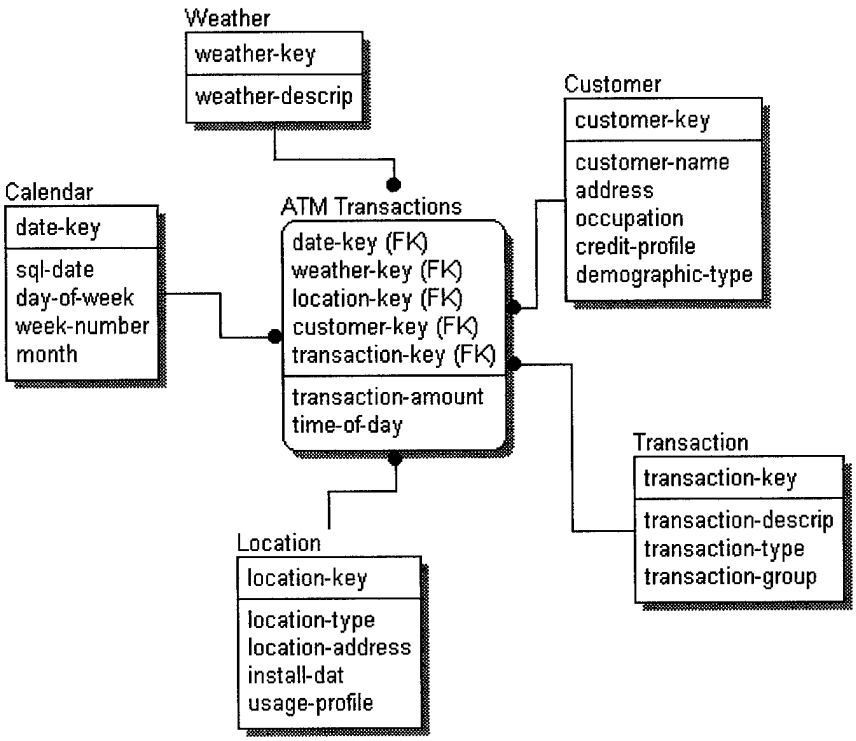

Figure 5 - Star Schema (14)

3.3.3 Facts, Dimensions, Attributes

Other design details include the facts, dimensions, and attributes to be stored in the warehouse. Facts are the representation of measurements of the business events. A typical fact for a weather reporting station might record the details of a weather observation noted at a given time. In a relational database, facts usually correlate with the transactions applied to the database. They could also be interpreted as summations of the changes to the database. Facts are the heart of the data warehouse and will occupy the most space.

In order for information to be extracted from the facts for analysis, the data is grouped and compared in various ways. The dimensions describe these groupings. Typical dimensions might include time, location, observer identity, etc.

Attributes are the individual detailed data items contained within the facts or dimensions. A location dimension would have attributes of the station identifier, station name, major command, climate type, and other identifying details. Facts have similar attributes. All of the facts, dimension, and attributes must be identified in the data model.

3.3.4 Granularity

An important property of the fact tables is their granularity. Granularity refers to the level of detail stored in the data warehouse. Choosing the granularity of the data is a critical issue. If it's too fine, the volume of data is increased and performance suffers. For instance, if a record is added to the database for every hour a forecast is recorded, it will require more storage than if a record is stored only when the forecast changes. If the granularity is too course, the user won't have access to the information he desires. Looking at the previous example, it may be more difficult to match the changes to the forecast with the hourly observations than it would be if the forecast were stored in hourly increments. Granularity not only has a profound impact on the size of the data warehouse, but the dimensions are also identified based on the grain of the fact tables.

3.3.5 Partitioning

One method of increasing the performance of a data warehouse is called partitioning. By breaking the data up into small physical units, access times can be decreased. Data can

be partitioned based on any combination of a number of different criteria including time, location, organizational unit, customer type, etc. Any of the dimensions of the warehouse are candidates for these criteria.

3.3.6 Identifying Data Sources

In any given business, the same data may be utilized by a number of different applications. Consequently, it is most likely stored in a number of separate locations or databases and on various platforms. The development team must determine which sources should be used to populate the data warehouse based on accuracy, completeness, timeliness, closeness to original source (or least distortion), and structural similarity to the data warehouse. This seemingly simple task can have great political ramifications. Not only must the correctness of the data source be considered, but data ownership should be taken into account as well. Confidence in the data set and future accessibility of the data source may depend on how well that source is maintained.

3.3.7 Data Integration

One of the challenges with disparate data sources is how to integrate the same data from multiple sources. Data items from separate sources may be related and must be merged or may look similar but have different meanings. It's also possible for the same data item to be stored in different formats for separate applications. The same data stored in separate databases may not match. The decision of which data source is correct or how to represent both *correct* sets of information in the data warehouse must be made based on

commonly agreed upon business rules. Conflicts such as these must be identified and resolved before the integrity of the data warehouse can be trusted.

3.3.8 Slowly Changing Dimensions

Another issue that must be resolved is that of slowly changing dimensions. This is data, such as a weather category definition, which may change slowly over time. Is it important to keep track of how and when the data changed, or will the most recent version suffice? The answer may change depending on the historical significance of the data and on the users' requirements. Therefore, this question must be answered for each piece of data that changes slowly over time.

3.3.9 Aggregation

Aggregation is a valuable technique that can greatly increase the performance of a data warehouse. An aggregation table contains summations of the fact table based on a particular set of dimensions. The initial set of aggregate tables is an anticipation of the users needs and questions. For instance, "How many times did any station in a specific major command (MAJCOM) amend a forecast in the first 6 hours?" is a typical question Air Force Weather might ask. It's efficient to pre-calculate the answers to the questions commonly asked by users and to store the results in aggregate tables. Some or all of the potential calculations are performed in advance so that *ad hoc* user queries can be accomplished quickly. The performance of the data warehouse is increased as the calculations are only performed when loading the data as opposed to each time the question is asked.

3.3.10 Metadata

Metadata is another critical issue that presents a significant obstacle because the collection and maintenance of metadata is predominantly a manual process. Metadata is information about the data stored in the data warehouse. It could include information such as the source of the data, a description of the field, allowable entries, the meaning of any codes used, the transformations or formulae applied, confidence levels for those formulae, etc. Across the literature, the list is extensive. The purpose of metadata is two-fold. Some metadata serves as technical reference for the data warehouse development and maintenance team. Other metadata enables the user to understand the context and meaning of the stored data.

3.4 Design the Warehouse

The data model described in paragraph 3.3 is a significant part of the data warehouse design. There are also a number of other factors that must be integrated into the design. The major components and how they relate to each other must now be identified. Figure 6 illustrates one such design.

Accompanying this step is the choice of hardware and software topologies. Selecting hardware and software is not an easy task due to the proliferation of available choices. To illustrate, DM Review surveyed over 690 data warehouse and business intelligence vendors before identifying their top 100 choices for the year (5:38). The overall architecture must also be determined. The decision to use a desktop, client-server, shared file, distributed, or network architecture is influenced more by the business' organization, integration with

existing systems, and user preferences than by any specific data warehouse driven requirements. As such, it is outside the scope of this thesis. However, Captain Jim Douglas' recent research offers a distributed solution for weather metrics (8).

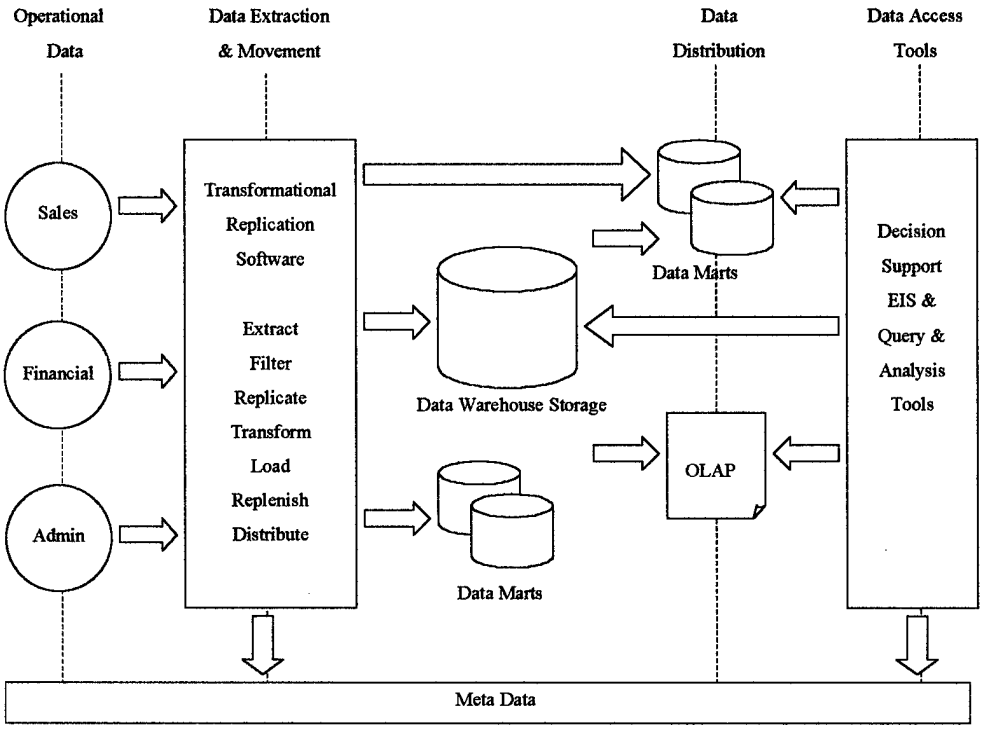

Figure 6 - Data Warehouse Architecture Diagram (9)

3.5 Data Retrieval (extract & load)

Retrieving data from its source is an on-going process in the data warehouse. Periodically the data needs to be extracted from the operational system and loaded into the data warehouse. In the process, it may have to be corrected and transformed into the proper format. Data maintenance and retrieval should not interfere with the operation of the originating data store.

3.5.1 Extract the Data

Extracting data from the originating data store is the first step of data retrieval. Because the data sources may run on a variety of platforms using diverse file formats, the data may have to be copied to a format that the data warehouse tools can recognize. Another issue concerns how much data is extracted. Because data retrieval is accomplished periodically, some of the data in the source system might have already been loaded into the data warehouse. Duplication of this data could have serious consequences, such as distorting the results of any statistical analysis, and should be avoided.

3.5.2 Correct and Integrate the Data

Errors, omissions, irregularities, and inaccuracies are commonplace in disparate databases and in some operational databases in general. It is important to correct these anomalies before the data is loaded into the data warehouse. Examples of errors include reported dew points higher than reported temperatures and winds reported as blowing from 700 degrees. Check for legal values and format. Source applications may not have edited their input values. The intent is to create a data source the users can access with confidence.

3.5.3 Transform (Refine) the Data

Data transformations restructure, reformat, or modify the data to make it ready for use in the data warehouse. This involves looking up descriptions for codes and transforming them into the format used in the data warehouse, separating a single field into multiple target fields, and concatenating two source fields together. Transformation also

includes calculating aggregations by summing or averaging the data over the appropriate dimensions.

3.5.4 Update Frequency/Urgency

How often should data retrieval be accomplished? Data warehouse users are concerned with how current the data is. Operational users need to limit the impact of the retrieval process on their systems. These two major areas of concern should be balanced in a way that makes the most sense for the business. Included in this decision is the volatility or *shelf-life* of the data in the operational data store. Data may only be maintained for a certain amount of time before it is purged from the source.

3.5.5 Populate the Warehouse

The final step of data retrieval is to load the data into the warehouse. The initial load will be more substantial than subsequent loads depending on how much historical data is available and whether or not the user is interested in that data. Subsequent loads may be simply incremental additions of new data or updates to existing data. In either case, aggregations and slowly changing dimensions must also be updated at the same time.

3.6 Data Analysis

The advantages of a data warehouse can't be exploited unless there is an efficient method of extracting and analyzing the data. Traditional query languages have limited functionality when analyzing data. Standard Structured Query Language (SQL) doesn't support operations such as ratios, time calculations, statistics, and ranking. Custom

formulae and algorithms aren't supported by SQL either. Data analysis tools require this functionality.

3.6.1 On-line Analytical Processing (OLAP)

It is important to distinguish the capabilities of a data mart from those of OLAP. A data mart stores and manages data. OLAP transforms that data into strategic information. OLAP tools provide the capability to examine information at various levels of aggregation across different dimensions in a very flexible and intuitive interface. They provide the functionality for statistical and temporal data analysis that is missing in SQL. OLAP products differ from each other much more than other software tools such as relational databases and programming languages. Not all OLAP products will suit specific needs. Distinguishing characteristics include the amount of data handled efficiently, number of users, functionality, web capability, and ease of use. Understanding the operational needs of the users is a critical part of selecting an OLAP product.

3.6.2 Pre-calculated (Black Box) Reports

Standardized, periodic reports are the traditional method of analyzing data. Some users require standardized reports on a periodic basis. OLAP products do provide support for repeatable queries whose results can be viewed as a report. However, if a particular set of users should have restricted access to the information in the data mart, or would rather receive information in a more traditional manner, pre-calculated reports containing standardized information can be provided through applications written specifically for that purpose.

3.6.3 Data Mining

Data mining is an alternate method of analyzing the data in a data warehouse. Data mining is used to find patterns in the data without actually knowing where to look. It could be interpreted as fishing for information. Although there are tools that provide data mining capabilities, it is actually an extensive process unto itself and as such is outside the scope of this thesis.

3.7 Purge the Warehouse

The value of data wanes over time and at some point, data must be purged from the warehouse. The historical duration of the data warehouse is dependent upon both user requirements and on available space. Calculating the lifecycle of the data should be an active part of the design process.

There are a variety ways in which data can be purged including the following:

1. Delete the data in its entirety
2. Summarize the data and delete the detail
3. Archive the data to long term storage
4. Transfer the data from one level of the warehouse to another, such as from the data mart level to the data warehouse.

3.8 *Other Issues*

A number of other issues can have great impact on the data warehouse. These complications don't fit neatly into any one step of the process and may be on going issues throughout the entire process.

3.8.1 Complexity

One of the many challenges in data warehouse development is to design a system where the user's won't be overwhelmed by the complexity yet still have the flexibility to explore many options. Any complexity inherent in the data model should be transparent to end-users.

3.8.2 Political Will and Sponsorship

As in any extended project, management support has a big impact on the success of a data warehouse. A data warehouse merges information from data sources managed by organizations with potentially conflicting interests. Financial support can also make or break the project.

3.9 *Summary*

The steps presented in this chapter broadly overview a methodology for developing a data warehouse or data mart. The process of designing, constructing, and maintaining a data warehouse has many interleaving steps and decisions that must be performed iteratively for any hope of success. Inmon's statement that "Data warehouses are not built all at once. Instead, they are designed and populated one step at a time, and as such are

evolutionary, not revolutionary." applies to the data warehouse as a whole, not to the individual phases of development (12:43). Many of the issues are interdependent; this is a considerable part of the reason data warehousing is such a challenging field.

IV. Results and Analysis

On first consideration, development of a data warehouse or data mart appears to be the perfect solution for on-demand dynamic metrics production because it would provide immediate access to the underlying data. Unfortunately, the typical means of analyzing the data in a warehouse utilizes additive amounts such as quantities and dollar figures to compute common functions such as sums, averages, or counts. Therein lies the challenge of applying a data warehouse solution to the weather metrics problem. Much of the statistical analysis of weather metrics is based on measurements and predictions of weather phenomena and the differences in those measurements and predictions for the same time period. This chapter describes how these measurements and predictions are successfully utilized in a data mart to produce on-demand dynamic metrics despite the atypical data.

4.1 Problem Definition

The subject of interest to Air Force Weather and to this research is how well the weather forecasters have supported the Department of Defense community. AFI 15-114 describes observation and forecasting skills as the foundation for effective weather support to the warfighter (6:2). Measuring those skills in terms of operational effectiveness, technical performance and wartime forecasting proficiency forms the basis for the AFW metrics program. It is this metrics program which is the focus of this research and the scope of the resulting data mart.

4.2 Requirements

Although AFI 15-114 is the initial driver for requirements, it is noticeably vague about the process of collecting and analyzing weather metrics. It does require each unit to develop metrics that are operationally significant, focusing on how well the unit supports the warfighter. Metrics should be developed and measured at the work center level in enough detail to drive improvement. Higher levels of management simply require an aggregation of the metrics collected at the unit level. Any system developed in support of these requirements will require a significant amount of flexibility in order to support the varied needs of each unit. The specific details of how to collect and analyze weather metrics will have to be discovered elsewhere.

The early versions of the TAFVER software provided limited statistical analysis of TAFs. In addition, the statistical results were continuously questioned since not all of the TAFs issued by field units were received for analysis. A conversation with the current AF CCC Chief Scientist revealed that TAFVER IV, the latest prototype version of TAFVER, was cancelled in 1998 when funding was cut for the program as opposed to any known inadequacies of the system (18). Therefore, we will proceed with the guarded assumption that the TAFVER IV requirements are still valid. Many of the following requirements are based on the TAFVER IV draft functional and statistical requirements (27).

Six topics to be considered when defining requirements were listed in paragraph 3.2. A discussion of each of these topics follows:

4.2.1 Subject Area

The problem definition narrowed the scope of the data mart to the Air Force Weather metrics program. This research further limits the subject area to the comparison of textual TAFs and the actual weather observations. Specific attention will be paid to the statistical analysis defined in the TAFVER IV requirements (27:8-13).

4.2.2 Atomic Level of Fact Detail

Assessing the quality of the weather forecasts depends on evaluating two sets of information: forecasts and observations of the weather. The statistical formulae currently used to calculate weather metrics are based on comparing the forecast to observations of the actual weather at specific times during the valid period of the forecast. These formulae are described in Appendix B. This analysis requires access to specific elements of each forecast and the associated observations.

Although the level of detail stored could be limited to the differences between the forecasts and observations, this would prevent future modifications to the way the data is analyzed. Any new statistical methods may then require major modifications to the data mart. Storing the differences but not the actual forecasts and observations has the advantage that the necessary storage space for the detail records can be reduced. However, this choice limits the potential of the data mart.

Categories, which are used extensively in the evaluation of the quality of the weather forecasts, are the complicating factor. Categories are defined intervals of values for each weather element. For example, one category for cloud ceiling might be greater than or

equal to 1500 feet and less than 3000 feet. Thresholds are the delineating values between the categories. The categories are different for each MAJCOM and for each weather element. For instance, Air Force Material Command may define three categories of wind speed at less than 25 knots, winds between 25 and 35 knots, and winds greater than 35 knots while Air Combat Command defines six categories of wind speed with thresholds at 10, 20, 30, 40, and 50 knots. Since the weather analysts should also be allowed to designate their own unique categories and the predefined categories may change over time based on the current needs of the MAJCOMs, it is necessary to maintain the original forecast and observation data in order to determine the current categories of the weather elements contained therein.

One of the goals of the data mart is to maintain flexibility. The ability to calculate metrics using formulae and statistical methods different from those currently defined and to assign categories as required by the current situation requires access to the original forecast and observation data. Therefore, the atomic level of detail will be the individual forecasts and observations.

4.2.3 Length of Fact Detail History

According to the TAFVER IV functional/statistical requirements, the data should be maintained on-line or near-line for the last five years and off-line for the previous six to ten years providing a total of ten years worth of data. Depending on available storage, a possible alternative to archival is to store the full ten years on-line. This would eliminate the overhead associated with restoring the archived data and ensure that it is readily available.

4.2.4 Required Business Dimensions

The weather analyst needs to be able to view results based on location, date, time of day, forecast group, and weather element. Specific divisions of the location dimension include the MAJCOM, squadron or hub, block station number, geographic region, aircraft type, weapon type, and climate type. Date intervals include daily, monthly, seasonally, quarterly, and annually. Combinations such as a group of stations or a period of consecutive days or months should also be able to be viewed. The time of day dimension consists of any combination of hours in the 24-hour period. The forecast group determines if the segments of the forecasts being verified are temporary, rapidly changing, or changing normally. Finally, the specific weather elements currently of interest are wind direction, wind speed, crosswinds, wind gusts, surface visibility, present weather, altimeter setting, and ceiling. Other weather elements may come of interest in the future; therefore, all weather elements recorded in the Aviation Routine Weather Report (METAR) format should be considered as candidates for analysis. (See Appendix A for the METAR format.)

4.2.5 Multidimensional Aggregation Requirements

The analyst may need to view the data based on any combination of the various dimensions. A typical request would be for results listed by station for a particular month, by weather element, and by forecast group. These results might then be broken down further into 6-hour intervals.

4.2.6 History Tables

History tables document slowly changing elements such as the names of MAJCOMs or squadron designators. In this application, tracking changes to the location data is probably unnecessary. The only changes to the location data that may be important to track are those that are a result of a significant re-organization to the unit where the block station number and International Civic Aeronautics Organization (ICAO) code don't also change as a result. In this case, the effective date of the change would be an indicator of changes in policy that may have affected the quality of the forecasts. This potential requirement has little impact on this research and will be disregarded.

The potential for changes to the categories provides a possible candidate for a history table. If the way the weather is forecast is dependent upon the categories in effect at the time of the forecast, it is important to identify changes to those categories. A forecast considered adequate at one point in time might have been inadequate if done under different circumstances such as those brought about by a change to the type of aircraft. This requirement will need to be substantiated with further research into the qualities that define a good forecast.

4.3 Data Model

4.3.1 Enterprise Data Model

Air Force Weather is in the process of developing their Reengineered Enterprise Infrastructure Program (REIP). This program will provide an enterprise-wide view of their data and the systems that manipulate it. Once this is complete, the data mart model

proposed herein should be compared to the enterprise data model to ensure the two models are consistent.

4.3.2 Relational Versus Multidimensional Versus Hybrid

A star schema was chosen as the data model for this research because this hybrid model is more understandable to the typical customer than either of the other models. Further considerations included the limited functionality of the relational model and the size limitations of the multi-dimensional model. Because statistical formulae are used in the data analysis, the relational model wasn't feasible. Standard structured query languages don't have the power to perform the required calculations. The multi-dimensional model has severe space restrictions that make it inadequate for the quantity of data to be analyzed.

4.3.3 Facts, Dimensions, and Attributes

As discussed in paragraph 4.2.2, the primary fact tables will contain forecasts and observations. A weather domain knowledge expert would have to be consulted to determine the exact attributes required. The attributes for forecasts and observations identified for the purpose of this research are the weather elements reported in the METAR format. Identified forecast attributes include the type of report (TAF, TAF AMD), forecast group (FM, TEMPO, PROB, BECMG), station identifier, issue date and time, valid period, wind direction, wind speed, maximum wind speed, wind variability from low to high, prevailing visibility, runway visual range, present weather, sky condition, high and low air temperatures, dew point temperature, altimeter setting, and remarks. Observation attributes include the following: type of report (METAR, SPECI), station identifier, date and time of

report, report modifier (AUTO, COR), wind direction, wind speed, maximum wind speed, wind variability from low to high, prevailing visibility, runway visual range, present weather, sky condition, air temperature, dew point temperature, altimeter setting, and remarks.

The dimension tables include date, time, and location. Attributes of the date dimension are the full date, day, julian day, month, year, week of year, quarter, and season. Attributes of the time dimension are the time, hour minute, and shift. Attributes of the location dimentison are the station identifier, block station number, station name, elevation, runway heading, MAJCOM, hub, geographic region, climate type, aircraft type, and weapon type. An additional attribute of the zulu time difference may be needed in order to accurately compare locations in separate time zones. The forecast group doesn't need to be a separate dimension as there aren't any attributes to be stored.

Finally, the additional category history table includes group name (typically MAJCOM), category, low threshold, and high threshold attributes.

4.3.4 Granularity

Because all metrics are based on the original forecast and observation data and the option for new types of metrics should remain available, the details of these forecasts and observations must be stored for future reference. Most of the data initially required for analysis will come from the aggregate tables. However, as the analysts discover new methods of evaluating the forecast, it may become necessary to store one or more forecast records for each observation rather than grouped by valid period. While this would be a

minor change to the data model, it would significantly increase the storage requirements. For now, the forecast records will be stored based on the forecast group and valid period.

4.3.5 Partitioning

The data mart should be partitioned not just because of storage restrictions and performance benefits, but also for security reasons. Allowing one MAJCOM or hub access to another's data should only be allowed upon justification of a special request. The easiest way to enforce this is to partition the data based on location. This could be accomplished by splitting the data sets between the separate hubs with access at the MAJCOM and HQ levels or by replicating the data at the higher levels. Each method has its own advantages and disadvantages, which are discussed in more detail by Capt Jim Douglas in his distributed engineering research (8).

The time dimension provides another opportunity for partitioning. As stated in paragraph 4.2.3, the original requirements called for the most recent five years of data to be stored on-line with the previous five years archived. This indicates that the older data will seldom be accessed and should be placed in separate partitions. The potential exists for even smaller partitions; however, usage patterns will have to be analyzed in order to determine effective time intervals for these partitions.

4.3.6 Identifying Data Sources

The Advanced Meteorological Information System (AMIS) and the National Weather Service web site (4) are the two current choices for forecast and observation data sources. Both sources provide the data as a single semi-formatted field in ASCII text format (see Figure 7 and Figure 8). The National Weather Service web site includes data from many other sources. Their database would have to be searched daily for applicable records, which would take significantly more time than extracting the information from AMIS. Since AMIS is not only readily accessible but it's also the original source of the applicable National Weather Service data and because the user has more control over the data, it's the more likely candidate for a data source.

```
TAF KOFF 130019Z RTD 130024 14008KT 9000 BR BKN009 OVC014 620144 QNH2997INS
   TEMPO 0306 4800 -FZRA BR SCT005 OVC008 690003 650802
   BECMG 0506 04012KT 3200 -SN BR SCT005 OVC008 620307 QNH2976INS
   BECMG 0809 04013KT 1600 SN BR SCT004 OVC008 610307 QNH2970INS
   BECMG 1718 02010G18KT 4800 -SN BR SCT004 OVC012 610208 510005 QNH2965INS
   BECMG 2021 36012G18KT 9000 BR OVC015 610208 510005 QNH2968INS TM01/18Z TM04/13Z
```

Figure 7 – Offutt AFB TAF (METAR format)

```
METAR KOFF 130555Z 06003KT 1 1/4SM -FZDZSN BR SCT003 BKN006 OVC010 M02/M02 A2985
RMK SLP119 60000 8/6// 9/8// 58008 IR11
```

Figure 8 – Offutt AFB Observation (METAR format)

Unfortunately neither one of these choices is ideal. In either case, the forecast or observation record will have to be parsed into its separate attributes. METAR and TAF decoder software provided by the Air Force Weather Agency will parse these text files. However, any records with errors are discarded in the process, leading to incomplete data.

Unless the missing data is identified, corrected, and input into the data mart, resulting data analysis is suspect and can't be trusted. Another issue with any parsing process is the inherent problem of ensuring the result of the parsing corresponds exactly to the original input as intended by the forecaster or observer. If the forecaster unthinkingly used the string 'fm' in the remarks field, the parser would interpret it as a rapidly changing forecast group instead of whatever was originally intended. There is also valuable data within the remarks of the METAR format, such as high and low temperatures, which can't be captured by the parser. Both AMIS and the National Weather Service are unsuitable as data sources not only because of the effort required to identify and correct erroneous records but also because the act of correcting those records will skew any analysis of their quality.

The critic advice system proposed by Capt Darryl Leon would be a significantly better data source (17). To begin, parsing will be unnecessary; the data will already be in a format that will easily map into the data mart. Syntax errors that would have caused a record to be rejected by the parser will no longer occur. This type of error includes values outside of natural limits, mistyped letters, and misplaced fields among others. In addition, the forecast or observation will have been validated before it is released from his system, ensuring that it is consistent with recently observed weather patterns. Most of the challenges inherent to data retrieval will be eliminated and the resulting data mart will be much more reliable. For the purpose of this research, Capt Leon's system will be utilized as the source for the forecast and observation data.

Sources for the dimensional tables also must be identified. The critic advice system will not only provide the data for the fact tables; it is also a valuable source for the category history and most of the location dimension data. Any attributes not contained in the critic advice system will have to be input manually. However, the long-term impact of this effort will be minimal as the data changes infrequently and the greater part of the missing data will only have to be input once. Generation and update of the data and time dimensions, which are standard to many data marts, can easily be automated.

4.3.7 Data Integration

From the perspective of the data mart, data integration is unnecessary since the observation and forecast tables are populated from a single data source. The necessary integration of the various sources of observation data is performed by Capt Leon's critic advice system. The forecast data originates with the critic advice system; therefore, integration of that data is unnecessary. By relying on the critic advice system as the definitive source of data, the essential task of data integration can be relinquished.

4.3.8 Slowly Changing Dimensions

As discussed in paragraph 4.2.6, categories are a prime candidate for a history table. The effective period of the categories with their thresholds and the name of the MAJCOM that defined them will be stored.

4.3.9 Aggregation

All of the statistical analysis is performed based on the differences between the forecast and the associated observations. Crucial factors are the differences between the categories and the actual values of the weather elements in the forecast and observations. Efficiency can be improved by matching forecasts to the appropriate observations and pre-calculating and storing these differences. Analysis can then be performed based on these pre-calculated values. Figure 9 illustrates the matching of observations with forecasts.

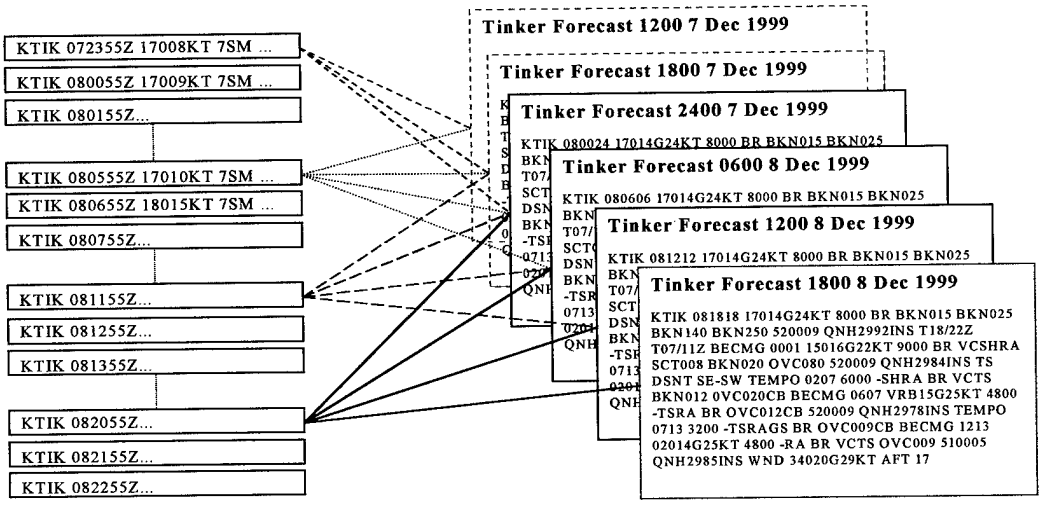

Figure 9 - Matching Observations to Forecasts

Depending on the data analysis software used, common combinations of analysis such as forecast quality for each MAJCOM by month, or for each station by season can also be aggregated as data is input into the data mart. These aggregations significantly increase the amount of storage required and should be carefully considered before implementation.

4.3.10 Metadata

Metadata will include a full data dictionary listing format and definitions of each field, along with a description of relationships between the tables and security information restricting access to the data mart. It should also include explanations of standard statistical analysis techniques and how the data stored in the data mart can be utilized in that analysis. Other interesting types of metadata are records of how the data mart has been utilized, such as logs of commonly occurring queries. The subject of metadata is a topic that bears further study and, as such, is outside the scope of this research.

4.4 Data Mart Design

Under the constraints discussed in the previous sections, and given highly standard data observation and forecast record formats, the initial design of a data mart is of minimum complexity. Distillation and implementation of all of the various components that were discussed in paragraph 4.3 leads to the data model shown in the following figures.

Each record of the forecast or observation tables (Figure 10) corresponds to a SurfaceCondition in Capt Leon's critic advice system. This in turn corresponds either to a single observation or to a segment of the TAF. The records of the main aggregate table (Figure 12) correspond to each matching pair of observations and forecasts.

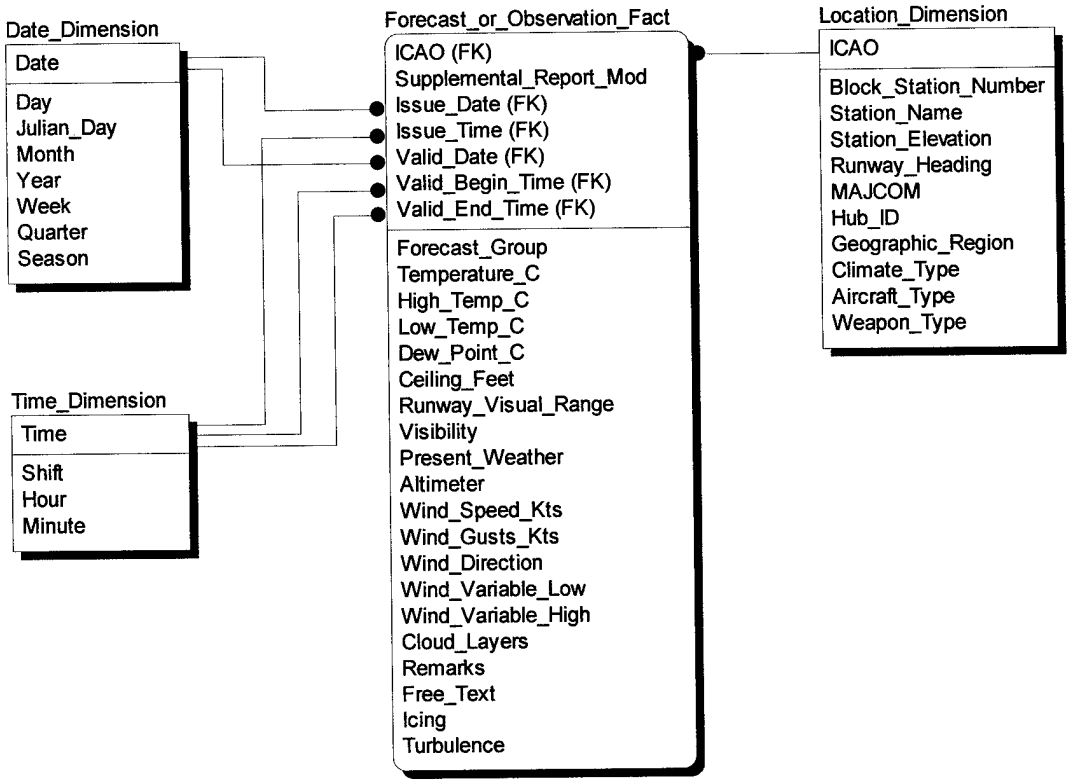

Figure 10 - Basic Data Model

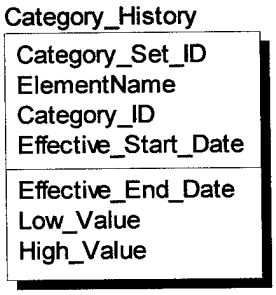

Figure 11 - Additional History Table

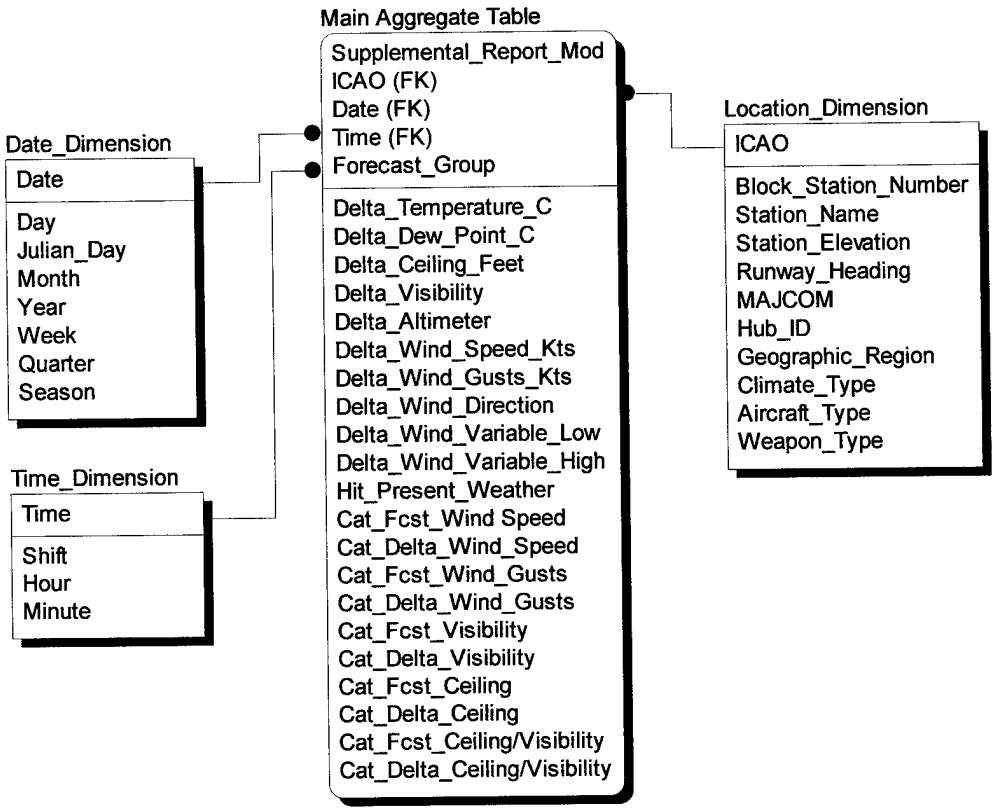

Figure 12 - Aggregate Data Model

The overall architecture as shown in Figure 13 is also quite simple. At each hub, forecasts and observations are entered into the critic advice system. The critic advice system stores forecast and observation records in an intermediate data store. The data is then extracted from the intermediate data store and transformed into the format required in the data mart. Calculations required to populate the aggregate table are part of this transformation. This data will be stored at the hub, and will be replicated at headquarters levels. This data replication concept is illustrated in Figure 14. As illustrated, regional data is collected at the regional weather squadrons or *hubs* and is then replicated at AFW. This replication includes applicable metadata. Weather observations can also be replicated at

the Air Force Combat Climatology Center (AFCCC) if needed. Each site requires a data retrieval engine (user interface) that will also perform statistical analysis.

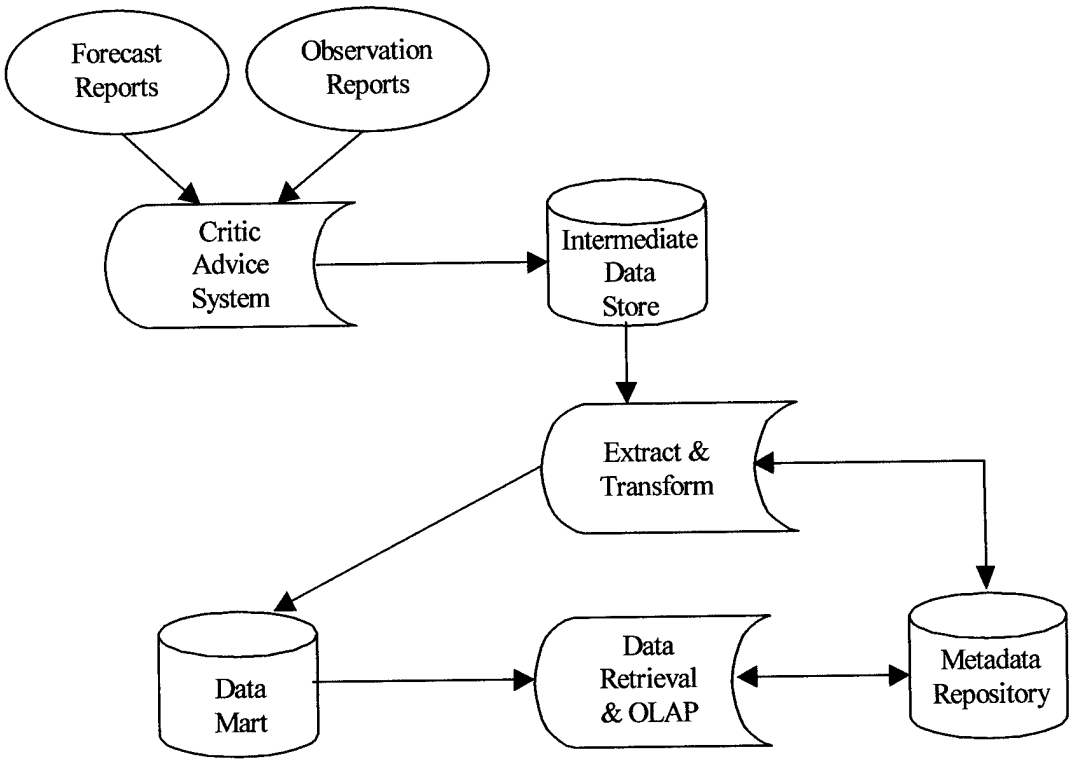

Figure 13 - Data Mart Architecture

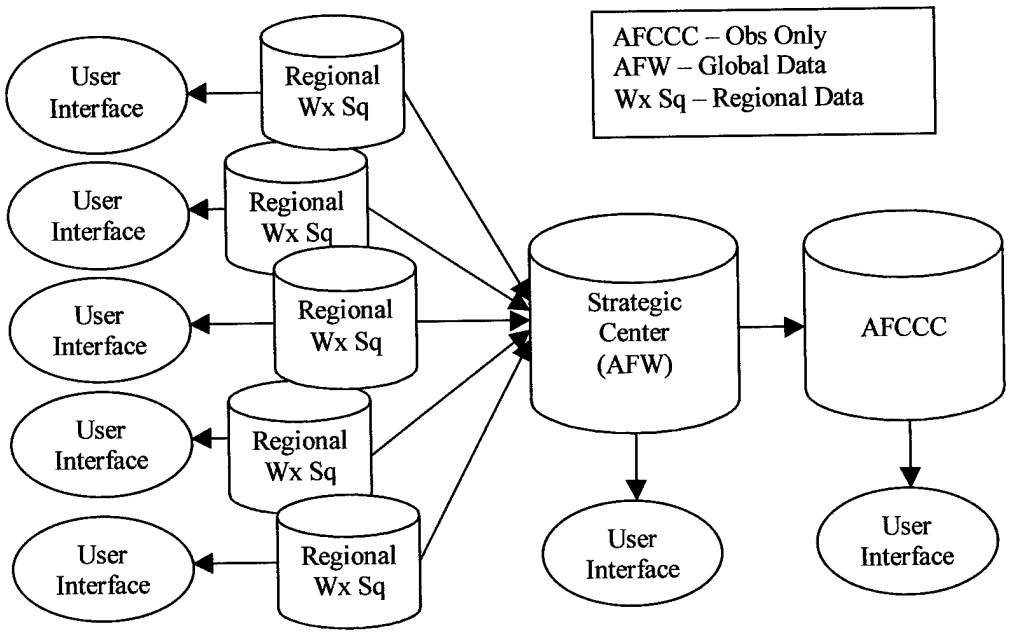

Figure 14 - Distributed Data Marts (Data Replication)

4.5 Data Retrieval

The principal purpose of the data retrieval process is to periodically insert new data into the forecast and observation fact tables. Challenges inherent to data retrieval are the primary reason that Capt Leon's critic advice system was chosen over AMIS as the source for the forecast and observation data. The problems that would have been encountered with AMIS include complications due to the parsing process, missing and duplicate records, disparate units of measurement, and values outside the natural range of an weather element. On the other hand, data retrieval from the critic advice system will be relatively straightforward.

4.5.1 Extract the Data

An intermediate data store is needed to capture the forecast and observation data as it is output from the critic advice system. Once this data store is established, the data can be extracted on a periodic basis. If the TAFVER IV requirements were referenced, this process could be performed monthly. However, AFW desires daily access to weather metric information (10). Performance of the extraction process on a daily basis would enable timely analysis of the information represented in the data mart.

4.5.2 Correct and Integrate

Since a suitably trained (in the machine learning sense) critic advice system edits and corrects the data as it is input into the system, no further correction will be necessary.

4.5.3 Transform

Capt Leon's object-oriented data model easily maps to the data mart star schema. His WeatherReport object contains all of the data needed to compose a group of forecast or observation facts. Each SurfaceCondition in the WeatherConditionSet along with the identifying data from the WeatherReport object will be transformed into a record in the forecast or observation fact table. The applicable attributes of the WeatherReport and SurfaceCondition all have matching attributes in the fact tables. This transformation is straightforward and trouble-free.

The transformation of the fact data into the aggregation tables is more complicated. Each observation must be matched to all corresponding forecast records and the differences between numerical elements calculated. The criteria for comparing present weather is

unclear and must be clarified by a weather domain knowledge expert before implementation. For the purpose of this research, it will be recorded as a hit or miss. Cloud layers are not compared at this time.

4.5.4 Update Frequency/Urgency

As the data from Capt Leon's critic advice system can be captured in an intermediate storage location, daily updates can be accomplished without interfering with the operational system as discussed in paragraph 4.5.1.

4.5.5 Populate the Warehouse

Once the transformation routines that map the data from the critic advice system to the data mart are set up, the task of populating the data mart should be merely repetitive. Because an intermediate data store will be used, data extract and population of the warehouse won't interfere with access to the operational data.

Since population of the data mart is an atomic process, the data mart cannot be accessed while this task is being accomplished. By populating the data mart on a daily basis, fewer records would be processed in each cycle; therefore, increasing the frequency but minimizing the duration of data mart down-time.

4.6 Data Analysis

Without the ability to efficiently retrieve and analyze data, a data mart is no more than just another data store. The various data analysis methods are where a data mart

empowers the analysts and decision makers. These methods include, but aren't limited to, OLAP, pre-calculated reports, and data mining.

4.6.1 On-line Analytical Processing (OLAP)

According to the OLAP council, "OLAP server technology is the key to high performance analytical use of large databases" (20). OLAP products not only answer the questions "who?" and "what?" but also "why?" The TAFVER software would have identified that a particular station's forecast quality was substandard. However, there was no way to *drill-down* into the data to ascertain the specific reasons for those poor forecasts. OLAP provides this capability. Once a problem area is identified, the analyst can ask meaningful questions to delve into the problem area. For instance, the analyst may want to know if the problem was isolated to a single shift or if there was a seasonal pattern. Other stations within the same climatic zone might be experiencing similar problems. All of these questions could easily be addressed with an OLAP product.

Because OLAP products are so diverse, it is hard to find right product for a particular application. *The OLAP Report* is an unbiased source of product information (21). It not only provides product reviews and analysis, the associated web site also contains advice on how to select and implement an OLAP solution, business-oriented case studies, white papers, and other OLAP specific information. At this time, an annual subscription to *The OLAP Report* costs $1540 for a single user. Multi-user subscriptions are available at discounted prices.

This area could not be fully assessed because the detailed information necessary for analysis is only available to subscribers of *The OLAP Report*. Because the typical OLAP application is commerce oriented, in-depth research into the specific capabilities of OLAP technology should be accomplished before the data mart is implemented.

4.6.2 Pre-calculated Reports

OLAP products can provide repeatable analysis of time series data. However, not every individual requires on-demand access to the data mart. Customers such as AF/XOW, Director of Weather, and his/her staff currently review monthly summaries of specific information. These monthly reports can either be provided utilizing the OLAP products or through software specially developed for that purpose. Pre-calculated reports might also be beneficial for any customers who don't have sufficient knowledge of weather quality analysis to ask meaningful questions. The wing commander and director of operations for a flying squadron are examples of customers who would find pre-calculated reports invaluable.

Although pre-calculated reports can be valuable, they don't take advantage of the data mart's strengths and shouldn't be the only means of data retrieval. OLAP and data mining are both powerful tools offering greater flexibility; OLAP tools also offer on-demand access. Pre-calculated reports merely complement their capabilities.

4.6.3 Data Mining

At this point, a foray into data mining is premature as its capabilities are outside the requirements of the weather metrics program. As the data mart matures and the analysts

gain experience, data mining could be employed to uncover previously unidentified trends. These might include patterns of forecast quality that are dependent upon the severity of the weather being forecast combined with the time of day, or any other combination of parameters. By adding significant information about the forecaster to the data mart, data mining could discover trends dependent on the rank, age, or skill level of the forecaster. Subtle trends that would have gone unnoticed by an analyst can be revealed using data mining techniques. Once the data mart has been successfully implemented, further research into data mining should be considered.

4.7 Purging Data

Referencing the TAFVER IV requirements, data older than 10 years should be purged from the data mart. Data between 5 and 10 years old should be transferred to a secondary partition or off-line data store. Purging data on a monthly basis will keep the size of the database relatively constant once it is fully populated while still allowing access to the data in daily, weekly, or monthly increments.

4.8 Other Issues

4.8.1 Complexity and Future Enhancements

Because the scope of the data mart was restricted to collecting metrics on the quality of the TAFs, the complexity of the system was minimized. As more functionality is added, the data mart will become more complex and expand into a full-fledged data warehouse. Functionality that might be considered in the future includes the comparison of cloud layers

forecast to those observed. As this comparison was not defined in the TAFVER IV requirements, it was not included in this design. However, once the algorithms to perform this comparison are developed, the necessary data can be integrated into the data mart design.

Another area of potential enhancement is the addition of the error data collected by Capt Leon's critic advice system. This system records any errors made during the input of the forecast or observation records. This data could be analyzed to identify input error trends by the same data analysis tools utilized with the forecast metrics data mart. The results of this analysis could then be utilized to enhance training.

As previously illustrated in Figure 14, AFCCC may find that either the critic advice system or the data mart is a suitable source for observation records. In this case, the data retrieval and analysis tool would need the capability to answer a different type of query. Additional aggregation tables might also be included in the design to facilitate these queries.

This data mart was designed with TAFVER IV requirements in mind. As additional functionality is added, consideration should be given to separating the developing data warehouse into separate data marts. As long as these separate data marts remain consistent with the enterprise data model, the complexity, from the user's point of view, will continue to be minimal.

4.8.2 Political Will

The current AF/XOW, Director of Weather, is the driving force in the Air Force Weather metrics program. While this level of interest imparts a sense of priority to the program, there is much other work to be done in AFW's reengineering effort and its effect on centralized TAF generation and distribution before a full-scale metrics program can be implemented.

4.8.3 Sponsorship

A critical issue that has yet to be addressed is funding. Data marts are not inexpensive; full-scale data warehouses typically cost between $1 million and $10 million (3). These figures include both one-time costs, such as the purchase or upgrade of hardware and software, and recurring expenses, such as the maintenance and update of the data warehouse, end-user training, and data warehouse administration. Like any software development effort, the knowledge required to develop and maintain a data warehouse does not come cheaply. The cost of hiring consultants is also included in the figures listed above. Fortunately, data marts can be developed much less expensively. In 1996, the cost of a data mart ranged between $200,000 and $2 million and has continued to fall dramatically ever since (11:8). Unfortunately, vendors don't release cost information unless they are dealing with a serious customer. For this reason, a full cost analysis could not be performed.

4.8.4 Estimation of Storage Requirements

A rough approximation of storage requirements provides a basis for preliminary estimates of the hardware requirements. This approximation was calculated as follows. Any given forecast is associated with as many records in the data mart as it takes to record the predicted changes in the weather over a 24 hour period. Amended forecasts are also recorded in the same manner. Observations are recorded every hour as well as whenever there has been a significant change in the weather. Because the number of forecast and observation records per day per station is somewhat unpredictable, it is hard to develop precise space estimations. For the purpose of this research, an estimate of the number of forecast and observation records per day per station was made based on forecast and observation data provided by the National Weather Service (4). On the particular day the estimate was made, there were an average of 20 forecast records and 28 observation records per day per station. The estimate of the number of weather stations was set at 100 after the AFCCC Weather Station Catalog was searched for United States Air Force Bases (29). Depending on how much space is allowed for remarks and free text, each forecast or observation record requires at least 200 bytes of storage. Each record in the main aggregate table requires less than 100 bytes of storage. A simplifying assumption was made that each observation record 'matched' with four separate forecast records (see Figure 15).

Initial calculations for the estimate of storage requirements are shown below.

$$100\ stations \times (20\ forecast + 28\ obs)\ records / station\ per\ day \times 200\ bytes / record$$
$$= 960{,}000\ bytes / day$$

$$100\ stations \times (28\ obs \times 4\ forecast)\ records / station\ per\ day \times 100\ bytes / record$$
$$= 1{,}120{,}000\ bytes / day$$

Each year of forecast and observation data will require 350 mb of storage (960 kb/day x 365 days). Each year of aggregate data will require 409 mb of storage (1.12 mb/day x 365 days) totaling 759 mb per year. The space required by the dimensional and history tables is negligible in comparison. If 10 years of data is stored in one data mart, it will require approximately 8 gigabytes of storage. This is well within the technological capabilities of today's storage devices. When distributed to local data marts, each data mart will require less than 2 gigabytes of storage.

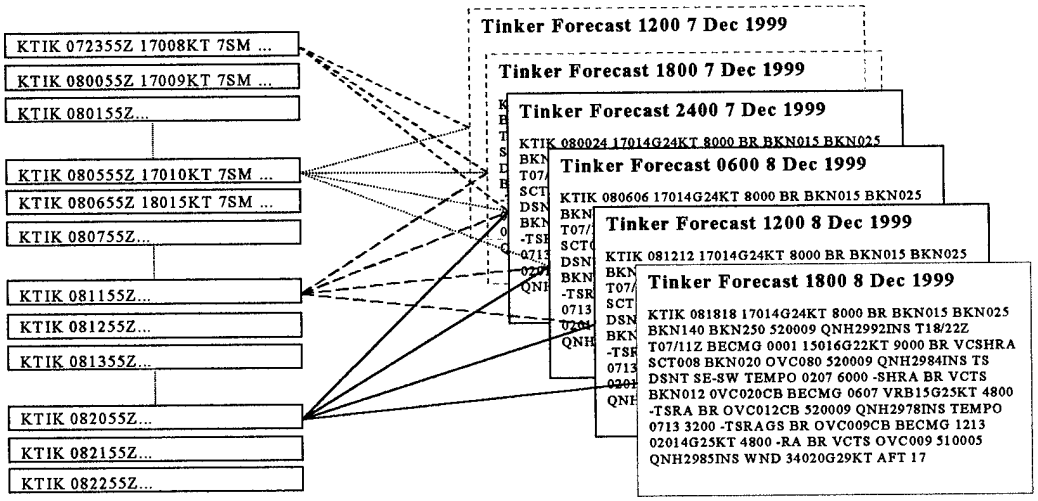

Figure 15 - Matching Observations to Forecasts

It should be understood that these initial estimates don't take into account the proprietary formats required by OLAP tools. Each OLAP tool handles data storage differently. Some perform calculations directly on data stored in a relational database every time information is requested. Others pre-calculate frequently used data and

permanently store the results in a multidimensional cube. Still others temporarily store calculated data on the chance that it will be accessed again relatively quickly. Indexing and metadata are other issues that will significantly impact data storage requirements. Precise storage requirements cannot be truly identified until the necessary data storage, retrieval, and analysis tools have been selected.

V. Conclusions

In retrospect, the actual research conducted was more related to defining a data warehouse development methodology than to its application to the weather metrics problem. Because standard data warehouse lifecycle models and development processes aren't yet firmly defined by the data warehousing community, a *best-of-class* methodology had to be developed before the problem of providing on-demand weather metrics could be addressed. Once the methodology was defined, it was applied with the following results:

5.1 Findings

Some method of edit checking and formatting the forecasts and observations as they are input must be implemented before a data mart will be feasible. The current METAR formatted forecast and observation records are unsuitable as data sources not only because of the effort required to identify and correct erroneous records but also because the act of correcting those records will skew any analysis of their quality. Even if every record could be parsed without error, the remarks field contains data which isn't accessible to the parser. There is no way to include this data in the data mart using the current record format. Capt Leon's critic advice system resolves these issues. His validation process not only eliminates syntax errors, it also identifies data that is inconsistent with the rest of the input record. The critic advice system also maintains each weather element in a separate field, eliminating the need for a parser or decoder. Implementation of his system or one like it should be seriously considered before a data mart is further contemplated.

The use of weather element categories overly complicates the design of the data mart. The concept of custom thresholds eliminates benefits associated storing pre-calculated data, which in turn impacts the efficiency of a data warehouse. If AFW can somehow find a way to identify good forecasts simply by the differences between the forecasts and observations, the implementation of a data mart would be much more straightforward.

Given an efficient data retrieval and analysis tool, a data mart can facilitate on-demand weather forecasting metrics production. Flexibility is the key benefit. With a data mart, analysis is not restricted to the pre-defined formulae; every manager can develop their own metrics as per AFI 15-114. Higher-levels of management can access summarized data without directing specialized data collection. When anomalies are identified, the data can be explored to find the reason for the irregularity. In short, a data mart introduces analysis possibilities that were previously inconceivable.

On the other hand, too much flexibility may be detrimental. Statistical analysis can easily be misinterpreted. If the managers don't understand the questions they are asking, the resulting information will be meaningless. Knowledge of statistical analysis techniques is necessary in order to produce valid results. User training on the development of metrics would have to accompany use of a data mart with separate pre-calculated reports produced for those users who don't require unrestricted access to the data mart..

Any data warehouse effort is an evolutionary process. The design presented herein duplicates the functionality of the outdated TAFVER software. Once implemented, it can be refined and the functionality will expand accordingly.

Although the data mart itself can be implemented, the ability to actually perform the required statistic analysis has yet to be demonstrated. Data retrieval and analysis tools are benchmarked against commercial enterprises using additive data. More research into these tools is necessary before a definitive result can be drawn.

5.2 *Recommendations*

To begin with, AFW should refine and implement Capt Leon's critic advice system. Without some method of ensuring the data is complete, analysis of the data can't be trusted. Beyond that, a critic advice system can provide actual improvements to the forecasting process versus simply reporting forecasting metrics.

In addition, further research into the capabilities of specific data retrieval and analysis tools should be performed to ensure the required statistical analysis can be accomplished. A subscription to *The OLAP Report* would provide an unbiased source of information on the multitude of available tools.

5.3 *Future Areas of Research*

5.3.1 OLAP

Data analysis and retrieval tools are benchmarked against commercial standards. Typical data warehouse vendors target the sales and financial industries. While the vendors claim their data analysis products efficiently perform statistical analysis, the question remains "Will the tools actually support the analysis required to produce forecast metrics or are they limited to standard commercial applications?"

5.3.2 Data Mining

Data mining has the potential to uncover previously unidentified trends as referenced in paragraph 4.6.3. Subtle trends that would have gone unnoticed by an analyst can be revealed using data mining techniques. Once the data mart has been successfully implemented, further research into data mining should be considered.

5.3.3 Forecast Metrics

In the course of this research, the question "What makes a good forecast?" arose repeatedly. The TAFVER IV requirements were based on research conducted prior to 1983 (7:2-6) and some as early as 1970 (28:3). In view of the advanced technology available today, the available research should be reviewed by a weather domain knowledge expert to ensure that the AFW forecast verification metrics are based on the most current statistical analysis techniques.

5.3.4 Data Warehouse Architecture

The literature on data warehousing is primarily concerned with development details. While the subject of how data warehousing fits into the system architecture as a whole is discussed, the data warehouse architecture is seldom mentioned. Research into how each of the elements of the data warehouse interrelate would be highly beneficial to the entire data warehousing community.

5.4 Summary

A data mart solution to the weather forecasting metrics problem has great potential. With a data mart, analysis is not restricted to the pre-defined formulae. Drill-down capabilities can be utilized to answer the question of why anomalies have occurred. However, Capt Leon's critic advice system is critical to successful implementation of a data mart. Without the data editing and formatting capabilities provided by the critic advice system, any resulting analysis is suspect. An efficient means of retrieving and analyzing the data is also critical as the potential benefits of a data mart cannot be realized without it.

Appendix A

National Weather Service
METAR/TAF Information (19)

KEY to AERODROME FORECAST (TAF) and AVIATION ROUTINE WEATHER REPORT (METAR)

TAF KPIT 091730Z 091818 15005KT 5SM HZ FEW020 WS010/31022KT
FM1930 30015G25KT 3SM SHRA OVC015 TEMPO 2022 1/2SM +TSRA
OVC008CB
FM0100 27008KT 5SM SHRA BKN020 0VC040 PROB40 0407 1SM -RA BR
FM1015 18005KT 6SM -SHRA OVC020 BECMG 1315 P6SM NSW SKC

METAR KPIT 091955Z COR 22015G25KT 3/4SM R28L/2600FT TSRA OVC010CB
18/16 A2992 RMK SLP045 T01820159

Forecast	Explanation	Report
TAF	Message type: **TAF**-routine or **TAF AMD**-amended forecast, **METAR**-hourly, **SPECI**-special or **TESTM**-non-commissioned ASOS report	**METAR**
KPIT	ICAO location indicator	**KPIT**
091730Z	Issuance time: ALL times in UTC **"Z"**, 2-digit date, 4-digit time	**091955Z**
091818	Valid period: 2-digit date, 2-digit beginning, 2-digit ending times	
	In U.S. **METAR**: **COR**rected ob; or **AUTO**mated ob for automated report with no human intervention; omitted when observer logs on	**COR**
15005KT	Wind: 3 digit true-north direction, nearest 10 degrees (or **VaRiaBle**); next 2-3 digits for speed and unit, **KT** (KMH or MPS); as needed, **G**ust and maximum speed; 00000KT for calm; for **METAR**, if direction varies 60 degrees or more, **V**ariability appended, e.g. 180**V**260	**22015G25KT**
5SM	Prevailing visibility: in U.S., **S**tatute **M**iles & fractions; above 6 miles in **TAF Plus6SM**. (Or, 4-digit minimum visibility in meters and as required, lowest value with direction)	**3/4SM**
	Runway Visual Range: **R**; 2-digit runway designator **L**eft, **C**enter, or **R**ight as needed; "/"; **M**inus or **P**lus in U.S, 4-digit value, **FeeT** in U.S. (usually meters elsewhere); 4-digit value **V**ariability 4-digit value (and tendency **D**own, **U**p or **N**o change)	**R28L/2600FT**
HZ	Significant present, forecast and recent weather: see table (below)	**TSRA**

FEW020	Cloud amount, height and type: **Sky Clear** 0/8, **FEW** >0/8-2/8, **SCaTtered** 3/8-4/8, **BroKeN** 5/8-7/8, **OVerCast** 8/8; 3-digit height in hundreds of ft; **T**owering **CU**mulus or **C**umulonim**B**us in **METAR**; in **TAF**, only **CB**. Vertical Visibility for obscured sky and height "VV004". More than 1 layer may be reported or forecast. In automated **METAR** reports only, **CleaR** for "clear below 12,000 feet"	**OVC010CB**
	Temperature: degrees Celsius; first 2 digits, temperature "/" last 2 digits, dew-point temperature; **M**inus for below zero, e.g., M06	**18/16**
	Altimeter setting: indicator and 4 digits; in U.S., **A**-inches and hundredths; (**Q**-hectoPascals, e.g. Q1013)	**A2992**
WS010/31022KT	In U.S. **TAF**, non-convective low-level (<=2,000 ft)**W**ind **S**hear; 3-digit height (hundreds of ft); "/", 3-digit wind direction and 2-3 digit wind speed above the indicated height, and unit, **KT**	
	In **METAR**, **ReMarK** indicator & remarks. For example:**S**ea-**L**evel **P**ressure in hectoPascals & tenths, as shown: 1004.5 hPa; **T**emp/dew-point in tenths °C, as shown: temp 18.2°C, dew-point 15.9°C	**RMK SLP045 T01820159**
FM1930	**FroM** and 2-digit hour and 2-digit minute **beginning** time: indicates significant change. Each FM starts on a new line, indented 5 spaces.	
TEMPO 2022	**TEMPO**rary: changes expected for < 1 hour and in total, < half of 2-digit hour **beginning** and 2-digit hour **ending** time period	
PROB40 0407	**PROB**ability and 2-digit percent (30 or 40): probable condition during 2-digit hour **beginning** and 2-digit hour **ending** time period	
BECMG 1315	**BECoMinG**: change expected during 2-digit hour **beginning** and 2-digit hour **ending** time period	

Table of Significant Present, Forecast and Recent Weather - Grouped in categories and used in the order listed below; or as needed in TAF, No Significant Weather.

QUALIFIER

Intensity or Proximity

- Light "no sign" Moderate + Heavy

VC Vicinity: but not at aerodrome;
in U.S. **METAR**, between 5 and 10SM of the point(s) of observation;
in U.S. **TAF**, 5 to 10SM from center of runway complex
(elsewhere within 8000m)

Descriptor

MI Shallow	BC Patches	PR Partial	TS Thunderstorm
BL Blowing	SH Showers	DR Drifting	FZ Freezing

WEATHER PHENOMENA

Precipitation

DZ Drizzle	RA Rain	SN Snow	SG Snow grains
IC Ice crystals	PE Ice pellets	GR Hail	GS Small hail/snow pellets

UP Unknown precipitation in automated observations

Obscuration

BR Mist(>= 5/8SM)	FG Fog(< 5/8SM)	FU Smoke	VA Volcanic Ash
SA Sand	HZ Haze	PY Spray	DU Widespread dust

Other

SQ Squall	SS Sandstorm	DS Duststorm	PO Well developed dust/sand whirls
FC Funnel cloud	+FC tornado/waterspout		

- Explanations in parentheses "( )" indicate different worldwide practices.
- Ceiling is not specified; defined as the lowest broken or overcast layer, or the vertical visibility.
- NWS **TAFs** exclude turbulence, icing & temperature forecasts; NWS **METARs** exclude trend fcsts
- Although not used in US, **C**eiling **A**nd **V**isibility **OK** replaces visibility, weather and clouds if: visibility >=10km; no cloud below 5000 ft (1500m) or below the highest minimum sector altitude, whichever is greater and no CB; and no precipitation, TS, DS, SS, MIFG, DRDU, DRSA, or DRSN.

March 1996
UNITED STATES DEPARTMENT OF COMMERCE
National Oceanic and Atmospheric Administration - National Weather Service
NOAA/PA 96052

Page Author: A.W. Jarvi, NWS Office of Systems Operations

Appendix B

Statistical Formulas

The following information was extracted from the TAFVER IV functional and statistical requirements (27:8-13)

a. Categorical Skill Scores and Statistics.

TAFVER III is to produce categorical skill scores and verification statistics according to the formulas described below. The variables "A", "B", "C", and "D" refer to values of the elements of the 2x2 matrices defined below. A typical 2x2 forecast verification matrix looks like this:

Forecast Weather Element to Verify: Cloud Ceiling, Category A (Cloud Ceiling < 200 feet)

		Forecast?		
		Yes	**No**	**SUM**
Observed?	Yes	**A**	**B**	**xx**
Observed?	No	**C**	**D**	**xx**
SUM		**xx**	**xx**	**xxx**

Element	What it means
A	Number of times that a category A cloud ceiling was forecast to occur (forecast = Yes) and then actually observed (observed = Yes) (regarded as a forecast "hit")
B	Number of times that a category A cloud ceiling was *not* forecast to occur (forecast = No) but then actually observed (observed = Yes) (regarded as a forecast "miss")
C	Number of times that a category A cloud ceiling was forecast to occur (forecast = Yes) but was not actually observed (observed = No) (regarded as a forecast "miss")
D	Number of times that a category A cloud ceiling was not forecast to occur (forecast = No) and was not observed (observed = No) (regarded as a forecast "hit")

1) Heidke Skill Score (HSS) is the ratio of occurrences versus non-occurrences; value ranges from -l to +1 is:

$$HSS = \frac{2(AD - BC)}{(A+B)(B+D)+(A+C)(C+D)}$$

2). Hanssen and Kuiper's Discriminant (or V Discriminant) ranges from -l to +1:

$$V = \frac{AD - BC}{(A+B)(C+D)}$$

3) Percent Correct. This is the number of correctly forecast events divided by the total, sometimes referred to as *forecast efficiency*:

$$PercentCorrect = \frac{A+D}{A+B+C+D}$$

4) Capability is the number of correct forecasts divided by the number of observed occurrences of the event.

a) Capability for "Yes" Forecasts. In this case, the capability shows the number of correctly classified forecasts over the total number of observed occurrences for a given category.

$$Capability = \frac{A}{A+B}$$

b) Capability for "No" Forecasts is calculated as follows:

$$Capability = \frac{D}{C+D}$$

5) Reliability. This is the number of correct forecasts divided by the number of forecasts issued for the event.

a) Reliability for "Yes" Forecasts:

$$\text{Re}\,liability = \frac{A}{A+C}$$

b) for "No" Forecasts is calculated as follows:

$$\text{Re}\,liability = \frac{D}{B+D}$$

6) Correlation coefficient (r). The correlation coefficient (ranging from -1 to +1) provides the linear correlation between observed and forecast events:

$$r = \frac{(AD - BC)}{(A+B)(A+C)(C+D)(B+D)}$$

7) Critical Success Index (CSI) is the number of correct predictions divided by the sum of the hits, false alarms, and missed forecasts.

$$CSI = \frac{A}{A+B+C}$$

8) AWS Skill Score (FSS) is the ratio of percent correct forecast less percent correct persistence to one minus the percent correct persistence.

$$FSS = \frac{\%Corr\_Fcst - \%Corr\_Persist}{1 - \%Corr\_Persist}$$

9) False Alarm Rate (FAR) is defined by:

1 - Reliability

A high false alarm rate is desirable in areas that are frequently threatened by severe weather and potentially widespread costly damage (i.e. areas threatened by hurricanes) to minimize lives and property loss.

(10) 4 x 4 Matrices. Matrices sized 4 x 4 are computed for the elements of ceiling and visibility only for both the basic forecast and persistence forecasts. These matrices look like:

Forecast Category

Observed Category	**A**	**B**	**C**	**D**	**Sums**
A					
B					
C					
D					
Sums					

In these 4x4 matrices, the variables "A", "B", "C", and "D" refer to categories of ceiling/visibility. These are not to be confused with the element values of the 2x2 tables defined above.

b. Non-Categorical Verification.

In addition to categorical verification statistics (yes/no occurrences of various threshold values), TAFVER III should be able to provide a root mean square error (RMSE) and a mean relative error (MRE). The RMSE is the average difference between the forecast and observed values for a given element. TAFVER should calculate RMSE and MRE for the following elements:

Ceiling Height (result in feet)

Wind Direction (result in degrees) (RMSE only)

Wind Speed and Gust Speed (results in knots)

Surface Visibility (result in meters)

Altimeter Setting (result in inches of mercury)

1) RMSE is calculated as follows for all the above elements:

$$RMSE = \sqrt{\frac{\sum (F_i - O_i)^2}{N}}$$

where:

Fi = the forecast condition at a given time

Oi = the observed condition at a given time

N = the number of forecasts.

2) ***MRE*** is calculated for all above elements except wind direction

$$MRE = \frac{\sqrt{\sum\left(\frac{F_i - O_i}{Oi}\right)^2}}{N}$$

where:

Fi - the forecast condition at a given time

Oi = the observed condition at a given time

N = the number of forecasts.

3) ***RMSE for Wind Direction*** is calculated as follows

$$RMSE = \sqrt{\frac{\sum (F_i - O_i)^2}{N}}$$

where :

If *(Fi - Oi) > 180*, Then

(Fi - Oi) = (360 -|Fi - Oi|)

4) RMSE for Visibility. The following rules are used to calculate RMSE for visibility forecasts :

a) If unrestricted visibility (coded "9999") is forecast and the observed visibility is 7 miles or greater, the error will be zero.

b) If unrestricted visibility is forecast and the observed visibility is less than 7 miles, RMSE is calculated as if the forecast was for 7 miles (e.g., if the forecast was for 9999 and 9000(6 miles) was observed, the error would be 1 mile (l,600 meters)).

5) RMSE for Ceilings. The following rules are used to calculate RMSE for ceiling forecasts:

a) RMSE for ceilings is calculated individually for each category to remove the bias that occurs due to errors in forecasting high-cloud ceilings. For example, a forecast for a 15,000-foot ceiling with an observed 20,000-foot ceiling would result in a 5,000-foot error, which is normally not operationally important.

b) RMSE is calculated only if a ceiling height is forecast. When no ceiling is forecast and a ceiling is observed (infinite error), no RMSE will be calculated.

Bibliography

[1] Advanced Weather Interactive Processing System (AWIPS) Program Description. National Weather Service. 17 Nov 1999. < http://www.nws.noaa.gov/msm/awips/descrip.htm >

[2] Air Force Weather Systems. System Requirements Document (SRD) for the Reengineered Air Force Weather Weapon System (AFWWS) Draft. Release 2.0. 30 December 1998

[3] Crofts, Steven. "Data Warehouse Consulting – Where Are We Today? A Research Summary," Data Warehousing: What Works, Volume 6: 32. 1998.

[4] Darling, Allan. National Weather Service Internet Weather Source. National Weather Service. 12 Feb 2000. < http://weather.noaa.gov/weather/coded.html >

[5] "The 1999 Data Warehouse 100." DM Review. December 1999: 38.

[6] Department of the Air Force. Weather Support Evaluation. AFI 15-114. Washington: HQ USAF, 19 January 1994.

[7] Donahue, Capt Christopher A. TAFVER II Users Manual. USAFETAC/TN-93-003. Scott AFB: USAF Environmental Technical Applications Center, May 1993.

[8] Douglas, James. Distributed Object System Engineering for Terminal Aerodrome Forecast Validation and Metrics Processing. Unpublished MS thesis, AFIT/GCS/ENG/00M-07. School of Engineering, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, March 2000.

[9] Drummond, Colin and Keegan, John. "Leeding Logistics Provider Leverages the Power of Data for Competitive Advantage," Data Warehousing: What Works, Volume 6: 32. 1998.

[10] Estes, Lt Col Frank. "FW: Thesis Project." E-mail to Col Ladwig and Key. 19 Feb 1999.

[11] Hackney, Douglas. Understanding and Implementing Successful Data Marts. Reading: Addison-Wesley; 1997.

[12] Inmon, W.H. Building the Data Warehouse. New York: Wiley, 1996.

[13] Jamison, Sherwin W. TAF Verification. AFGWC/TN-86/002. Offutt AFB: Air Force Ghlobal Weather Central, November 1986. (AD-A177931)

[14] Kimball, Ralph. "Help for Dimensional Modeling" DBMS Online. August 1998. 21 Feb 2000. < http://www.dbmsmag.com/9808d05.html >

[15] -----. The Data Warehouse Toolkit. New York: Wiley, 1996.

[16] -----. Ralph Kimball Associates: Who We Are. 25 Oct 1999 < http://www.rkimball.com/whoweare.htm >

[17] Leon, Darryl. An Intelligent User Interface to Support Air Force Weather Product Generation and Automated Metrics. Unpublished MS thesis, AFIT/GCS/ENG/00M-15. School of Engineering, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, March 2000.

[18] Lowther, Lt Col Ronald. Personal interview. 8 Dec 1999.

[19] METAR/TAF Information. National Weather Service. November 1998. 21 Feb 2000 < http://www.nws.noaa.gov/oso/oso1/oso12/document/guide.shtml >

[20] OLAP Council. Home page. 16 Feb 2000 < http://www.olapcouncil.org >

[21] Pendse,Nigel and Creeth, Richard. The OLAP Report. 16 Feb 2000 < http://www.olapreport.com >

[22] Relational OLAP: Expectations and Reality. Arbor Software. 19 Jan 2000 http://warehouse.chime-net.org/software/cisdss/rolapc0.htm

[23] Schmidt, SMSgt Don, ed. Air Force Weather Strategic Plan. Spec. issue of Observer 44.7 (September/October 1997): 1-39

[24] Silberschatz, Avi and Zdonik, Stan. "Database Systems – Breaking Out of the Box" SIGMOD Record. Volume 26 Number 3. September 1997: 36-50 20 Feb 2000 < http://www.acm.org/sigmod/record/issues/9709/index.html >

[25] Sumaria Systems Inc. TAFVER IV Systems Manual Version 1.0. Unpublished Manual for the Air Force Combat Climatology Center, Asheville NC. 4 Apr 1998

[26] Sybase. Data Warehouse Technical Guide. 27 Jan 2000 < http://www.sybase.com/products/dataware/techguide.html >

[27] Taylor, David A. Functional/Statistical Requirements. Draft, 23 Sep 1997. Electronic.

[28] Telfer, Ray T. and Webb, Randall C. Terminal Forecast Verification. Technical Note #70-2. Offutt AFB; Headquarters 3D Weather Wing, February 1970 (AD 707498).

[29] Weather Station Catalog. Version 1.0, IBM, diskette. Computer software. AFCCC/SYS, Scott AFB IL, March 1997.

Acronyms

AFCCC	Air Force Combat Climatology Center
AFW	Air Force Weather
AMIS	Advanced Meteorological Information System
AWDS	Automated Weather Distribution System
AWIPS	Advanced Weather Interactive Processing System
ICAO	International Civic Aeronautics Organization
MAJCOM	Major Command
METAR (roughly translated from French)	Aviation Routine Weather Report
NCEP	National Centers for Environmental Prediction
N-TFS	New Tactical Forecast System
OLAP	On-line Analytical Processing
REIP	Reengineered Enterprise Infrastructure Program
SQL	Structured Query Language
TAF	Terminal Aerodrome Forecast
TAFVER	Terminal Aerodrome Forecast Verification

Vita

Captain Meriellen C. Joga was born on 20 September 1960 in New London, New Hampshire. She graduated from Merrimack Valley High School, Penacook, New Hampshire in 1978. After enlisting in the Air Force in 1979 as a Munitions Systems Technician, she served tours at Nellis AFB, Nevada and Ramstein AFB, Germany. In 1990, she retrained into the communications and computer systems programming career field and was assigned to Kelly AFB, Texas as an information systems programmer. She was selected for the Airman Education and Commissioning Program (AECP) in 1992 and was reassigned to the University of Texas, San Antonio to complete a Bachelor of Science degree in Computer Science, where she graduated magna cum laude in May of 1994. After completing Officer Training School, her first commissioned assignment was to the Air Education and Training Command Computer Support Squadron (AETC CSS) at Randolph AFB, Texas, where she was the Officer in Charge of Computer Systems Maintenance. In 1996, she transferred to the Air Force Operational Test and Evaluation Center at Kirtland AFB, New Mexico and performed as the Electronics/Computer Safety Manager. She remained in New Mexico until August of 1998 when she embarked on her Air Force Institute of Technology adventure. Upon graduation, she will be assigned to Tinker AFB, Oklahoma.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	March 2000	Master's Thesis

4. TITLE AND SUBTITLE
DATA WAREHOUSE TECHNIQUES TO SUPPORT GLOBAL ON-DEMAND WEATHER FORECAST METRICS

5. FUNDING NUMBERS

6. AUTHOR(S)
Meriellen C. Joga, Captain, USAF

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Air Force Institute of Technology
Graduate School of Engineering and Management
2950 P Street, Building 640
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER
AFIT/GCS/ENG/00M-09

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Mr George Coleman, GM-14, DAFC
Acting Director, Plans and Programs Directorate
HQ AFWA/XP
106 Peacekeeper Dr STE 2N3
Offutt AFB NE 68113-4039 DSN 271-3585 Comm: (402) 294-3585

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
Prof Henry B. Potoczny, ENG, DSN: 785-6565 ext. 4282 Comm: (937) 255-6565 ext 4282

12a. DISTRIBUTION AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Air Force pilots and other operators make crucial mission planning decisions based on weather forecasts; therefore, the ability to forecast the weather accurately is a critical issue to Air Force Weather (AFW) and its customers. The goal of this research is to provide Air Force Weather with a methodology to automate statistical data analysis for the purpose of providing on-demand metrics. A data warehousing methodology is developed and applied to the weather metrics problem in order to present an option that will facilitate on-demand metrics. On-line analytical processing (OLAP) and data mining solutions are also discussed.

14. SUBJECT TERMS
Data Warehouse, Data Mart, Data Bases, OLAP, Data Management, Corporate Information Management, Weather, Forecasting, Meteorological Data

15. NUMBER OF PAGES
93

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reports. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow. It is important to ***stay within the lines*** to meet ***optical scanning requirements.***

Block 1. Agency Use Only *(Leave blank)*.

Block 2. Report Date. Full publication date including day, month, and year, if available (e.g. 1 Jan 88). Must cite at least the year.

Block 3. Type of Report and Dates Covered. State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

Block 4. Title and Subtitle. A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

Block 5. Funding Numbers. To include contract and grant numbers; may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C - Contract	**PR** - Project
G - Grant	**TA** - Task
PE - Program Element	**WU** - Work Unit Accession No.

Block 6. Author(s). Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

Block 7. Performing Organization Name(s) and Address(es). Self-explanatory.

Block 8. Performing Organization Report Number. Enter the unique alphanumeric report number(s) assigned by the organization performing the report.

Block 9. Sponsoring/Monitoring Agency Name(s) and Address(es). Self-explanatory.

Block 10. Sponsoring/Monitoring Agency Report Number. *(If known)*

Block 11. Supplementary Notes. Enter information not included elsewhere such as: Prepared in cooperation with....; Trans. of....; To be published in.... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

Block 12a. Distribution/Availability Statement. Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NOFORN, REL, ITAR).

- **DOD** - See DoDD 5230.24, "Distribution Statements on Technical Documents."
- **DOE** - See authorities.
- **NASA** - See Handbook NHB 2200.2.
- **NTIS** - Leave blank.

Block 12b. Distribution Code.

- **DOD** - Leave blank.
- **DOE** - Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports.
- **NASA** - Leave blank.
- **NTIS** - Leave blank.

Block 13. Abstract. Include a brief *(Maximum 200 words)* factual summary of the most significant information contained in the report.

Block 14. Subject Terms. Keywords or phrases identifying major subjects in the report.

Block 15. Number of Pages. Enter the total number of pages.

Block 16. Price Code. Enter appropriate price code *(NTIS only)*.

Blocks 17. - 19. Security Classifications. Self-explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contains classified information, stamp classification on the top and bottom of the page.

Block 20. Limitation of Abstract. This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.