Statistics Warehouse for Intelligent Networks

Applying data warehouse technology in the telecommunications domain

J. Schabauer, H. Portschy, R. Siebenhofer, A. Holzinger

SIEMENS AG AUSTRIA

Program and System Engineering

Telecommunications Management Networks - Intelligent Networks Department

Gudrunstrasse 11, 1100 Vienna, Austria

e-mail: {johann.schabauer, hans.portschy, rudolf.siebenhofer, andrea.holzinger}@siemens.at

Key words: Intelligent Network, Data Warehouse, Decision Support, Database

Abstract: Network operators and service providers are faced with technology and marketplace changes. Enhanced decision-supporting capabilities are needed to respond quickly to these changes. In this paper we present a system which applies a data warehouse approach for collecting and analysing data from an Intelligent Network to support business decision support systems. The system continuously gathers data from disparate systems in different formats, transforms the data into a consistent format and loads it into a database. Sophisticated and flexible analysis report facilities are provided, which are important for network planning, system operation and marketing.

1. INTRODUCTION

The telecommunications market is becoming increasingly competitive. Service providers and network operators are facing dramatic changes in the way they do business. The model of the past, one provider or just a handful of competitors in each country, is being replaced by a model shaped by hundreds of competitors competing for global presence. Under these conditions, the customer becomes the central focus of a provider's activities, and the ability to react quickly to market trends and to tailor products and services to individual customers is more critical than ever. Providers need to

The original version of this chapter was revised: The copyright line was incorrect. This has been corrected. The Erratum to this chapter is available at DOI: [10.1007/978-0-387-35522-1_37](https://doi.org/10.1007/978-0-387-35522-1_37)

have a complete understanding not only of their technologies, networks and offered services but also of what their customer usage behavior and desires are for the network operation, system planning and marketing.

The system we have developed collects and analyzes information from an Intelligent Network and presents this information in an understandable and flexible way to end users through a user-friendly Web-interface. These analysis reports and evaluations may then support a variety of decision analysis functions as well as strategic operational functions.

2. BACKGROUND

This section gives a short overview of the Intelligent Network (IN) and the application domain.

The IN essentially separates service control functions from service switching functions. Typically, both types of functions are implemented in different physical equipment. Due to this separation, it is possible to introduce new services into the network without having to change the functionality of the switches in the network. *Figure 1* shows the IN architecture.

Figure 1. IN architecture

Service switching points (SSPs) are stored program control switches that interface to the SS7 signalling network. The SSP embodies the call control function (CCF) and service switching function (SSF) entities. The SSF recognizes IN service calls and routes the appropriate queries to the service control function (SCF) that resides in a service control point (SCP) via the

SS7 network. SCP functions are used by the SSP to process calls. The SCP is a fault-tolerant, high-capacity, transaction-processing entity that provides call-handling information in response to SSP queries. The service management point (SMP) provides operation, administration, and maintenance functions for the IN.

All of these basic elements form the infrastructure of IN, which supports the notion of separating service-control functions from service switching-functions to realize more rapid-services development and deployment.

For each service usage SCP and SMP collect and log a lot of information. The SCP creates call detail records (CDRs), which contain all call related data including service number, calling party, called party, date, time, duration, release cause and so on. The SMP produces administration tickets, which include service number, date, time and service parameters. All these records and tickets contain a gold mine of information about the network behavior, the service behavior and the user behavior. The analysis and evaluation of this information is essential for the service providers to improve their quality of service and to understand the customers' needs and behaviors. It is also useful for the service subscribers and end-users to control and optimize their service usage.

3. CHALLENGES AND REQUIREMENTS

One of the characteristic challenges of the telecommunications business is managing enormous volumes of data. Not only do telecom systems generate huge numbers of call detail records (and others), but the records themselves are very large. Thus one of the most important and also critical success factors is to handle tremendous insertion rates, as well as to manage databases that will likely grow into multiterabyte range. Building a system that will not scale efficiently as data volumes grow can be one of the most frustrating (and expensive) experiences a company can endure.

Over the last years a popular approach for collecting and analyzing data and information is the use of data warehouse technology. Therefore, we have adopted a data warehouse approach to collect and aggregate statistical data in a continuous way (24 hours 365 days a year) for millions of calls per hour. We have termed our system Statistics Warehouse - collection and analysis of statistical data form an Intelligent Network using a data warehouse approach.

To support the call rate of IN services a data throughput of up to 50 million call detail records per hour is necessary. Call detail records have to be archived over 90 days. 50 million call detail records per hour over 90

days means that a storage volume up to 50 Terra Byte has to be managed. *Figure 2* illustrates the data rates and the data volumes.

Figure 2. Data rates and volumes

4. STATISTICS WAREHOUSE

This section describes the overall architecture of the system, the data model, the database configuration and the process of importing, transforming, and loading the data.

The overall architecture consists of three main parts: the mediation component, the warehouse component and the user interface component (see *Figure 3*).

Figure 3. Overall architecture

The mediation component periodically imports call detail records from the SCPs and administration tickets from the SMPs via file transfer. This data is cleaned and transformed into an internal uniform format for loading into the database. In addition, an export to other external systems (e.g. a billing system or another data warehouse) is also possible. This concentrated data is then loaded into the database to populate the tables.

The warehouse component consists of the database, a global data import function, data analysis and evaluation functions and an export mechanism to export filtered data and reports to other external systems. The global data import function is used to import IN-specific global data, such as announcements, origin areas and other network related data, from the IN management system. Backup and restore of the database is integrated with the OAM concept of the IN system.

The user interface component, which provides a user-friendly Web-interface, performs access control, invokes the evaluation and analysis functions and sends the reports to the Web-browsers.

The system has been developed using relational database technology, CORBA technology and Web Technology. Java has been chosen as the programming language for client and server components.

Due to the enormous amount of data and the data rate, the critical core components are the data model and the import and load mechanism. The following subsections describe these components in more detail.

4.1 Data model

For the data model we have adopted the standard practice of the star schema (see *Figure 4*). The star schema is a concept that provides superior data retrieval power to the data warehouse. The name "star schema" is derived from the appearance of the data model, with a large central table surrounded in a star formation by subordinate tables. The central table, known as the fact table, is surrounded by many dimension tables. This type of modelling is known as dimensional data warehousing. Fact tables contain measures that are used to perform analysis as well as the keys that link the dimensions. The dimension tables contain the attributes that describe the data components and provide the information to do comparative analysis.

Our model consists of the main fact table `raw_fact`, several dimension tables and the table `raw_rest`. All call detail records contain large fields like service number, origin, etc. which are identical in numerous records. These fields are not unique primary keys for all pre-defined report queries. Most queries affect a range of such keys, for example all origins of a district. To optimize the data access and minimize the data volume, the most-queried fields are stored in a separate fact table and ranges are stored in dimension tables. The rest of the raw data is stored in a so called rest table.

The `raw_fact` table contains the most-queried fields - such as number of calls, number of dialogs, number of reroutings, etc. - and the keys to the other tables. The `raw_rest` table contains all remaining fields. The dimension tables include amongst others time, origin, and service. Time is a dimensional table that holds date and time information. The origin dimension contains geographical information including country, district and city. The service dimension contains information about the IN services and numbers.

Figure 4. Data model

All raw data is located in `raw_fact` and `raw_rest` tables, which can contain up to 50 thousand million records. To manage such an amount of data, these logical tables are divided into smaller parts by partitioning of the tables by day. Partitioning by the day has the following advantages:

- Data for a specific day can be directly accessed.
- Loading of new data affects only one partition.
- Deletion of expired data is achieved by dropping the partition.
- Backup and restore can be done separately for each partition.

A daily partition can contain up to 500 million records, which is still a large number. Therefore each daily partition is divided into several sub partitions. This composite partitioning concept (see *Figure 5*), a combination of range (daily) for primary partitioning and hash for sub-partitioning divides one large table into manageable parts. These sub-partitions are distributed to several discs to enable good I/O balancing to achieve high performance data access.

		hash partitions by time (secondary)					
		H1	H2	H3	H4	...	Hn
range partitions by day (primary)	1.1.2000					...
	2.1.2000					...
	3.1.2000					...
	4.1.2000					...	
	...	...	...	...	...	...	...
	31.3.2000					...

Figure 5. Composite partitioning

The data model and the partitioning concept enable to handle a database with 50 thousand million records and a data volume of up to 50 TerraByte.

4.2 **Data Import and Transformation**

Data is imported from several different data sources; call detail records from the SCPs and administration tickets from the SMPs. This data is periodically imported via file transfer and stored in files. Due to the fact that the data is imported from several systems it usually has a different format. For each data source with a distinct data structure a corresponding interface module is provided. These interface modules perform data cleaning and transformation of data into a uniform structure. The output of this process are loadable files ready to be loaded into the database. *Figure 6* shows the architecture and the data flow of the import and transformation process.

Figure 6. Import and transformation process

To achieve high data throughput and to prevent loss of data, an import manager synchronizes and controls the parallel processing of the file transfers, the cleaning and transformation of imported data into uniform loadable files, and the hand-over of loadable files to the data loading process. For synchronization and error recovery without corruption or loss of data, each step in this process is treated as a transaction. The current status of each step and file is logged into a status table.

4.3 Data Loading

The data loading process has to support a load rate of up to 50 million records per hour. Therefore a sophisticated load mechanism has been implemented, where the data loading is done in several steps (see *Figure 7*):

- Step 1: Loading of the data into temporary tables to enable parallel processing of the bulk loader and to prevent I/O conflicts.
- Step 2: Adding foreign keys, creating indexes and new dimensions.
- Step 3: Insert-append of the temporary tables to the fact and rest_raw tables.

Figure 7. Data loading mechanism

This multi-step loading mechanism has the advantage that the large data tables are only affected in the short time of the insert-append and that all other processing steps are done with small, independent temporary tables.

5. ANALYSIS REPORTS

The system provides a set of pre-defined analysis reports. These reports can be customized to fulfil the different requirements and needs of the various users. This enables the user to perform any user-specific analysis and to access all of the data that he or she could possibly desire. Further evaluations can be done using Online Analysis Processing (OLAP) tools. The data model as well as the system architecture allow an easy integration with other systems and analysis programs. In addition raw data filtering and download of the filtered extract of raw data to the client workstation for further processing/analyzing is also supported.

The analysis reports are grouped into IN service-independent evaluations, which are based solely on call detail record fields available for all services, and IN service-specific evaluations, which are based on fields in call detail records, fraud records and confirmation records available only for certain services. All analysis reports are grouped and sorted by IN number, date and time.

The service-independent analysis allows the user to obtain the following reports:

- Traffic curve:
 - total number of calls, average call and set-up duration, number of successful calls sorted by call origin, SCP number or SSP number
- Analysis of successful calls:
 - number of calls sorted by set-up and call duration

- Analysis of rejected calls:
number of calls sorted by reject cause
- Structural analysis:
number of calls sorted by origin and/or destination location

The service-specific analysis allows the user to obtain the following reports:

- Analysis of voting result:
number of calls sorted by service embedded counters
- Fraud analysis:
number of events sorted by event type (e.g. wrong PIN, simultaneous call)
- Call charge analysis:
number of calls sorted by amount of charge
- Recharge analysis:
successful rechargings, unsuccessful rechargings, recharge amount, average recharge amount
- Outpayment analysis:
successful outpayments, unsuccessful outpayments, average amount of outpayment

6. WEB-BASED USER INTERFACE

All user interactions such as analysis reports, raw data filtering, system administration and user administration are done with a graphical user interface running in a standard Web-browser.

Analysis reports and raw data filtering are performed on demand. The user initiates an evaluation. This order is stored and processed by the warehouse component. The result is written to a file, which can be displayed on the browser or exported to an external system. A user management provides privacy of information and prevents unauthorized access and user actions. Access rights, roles and restrictions can be defined for each user by the system administrator.

7. CONCLUSIONS

This paper describes the Statistics Warehouse for Intelligent Networks, a Web-based system for the analysis of the behavior of IN service users. To fulfil the performance and flexibility requirements a data warehouse approach has been used. The Statistics Warehouse provides different kinds

of analysis reports and evaluations which are essential for decision support purposes in the area of network operation, system planning and marketing.

REFERENCES

- [1] Harry S. Singh, *Data Warehousing Concepts, Technologies, Implementations and Management*, Prentice Hall, 1998
- [2] Michael J. Corey, Michael Abbey, Ian Abramson, Ben Taub, *Oracle8 Data Warehousing*, Osborne/McGraw-Hill, 1998
- [3] Ralph Kimball, *The Data Warehouse Toolkit*, Wiley, 1996
- [4] Gregory F. Pfister, *In Search of Clusters*, Prentice Hall, 1998
- [5] Thomas J. Mowbray, Raphael C. Malveau, *CORBA Design Patterns*, Wiley, 1997
- [6] Andreas Vogel, Keith Duddy, *JAVA Programming with CORBA, Advanced Techniques for Building Distributed Applications*, Wiley, 1998
- [7] *Oracle Express Server - Delivering OLAP to the Enterprise*, An Oracle White Paper, January 1997