OPTIMAL DATA ARCHITECTURE FOR AN TELEMEDICINE ANALYTIC PLATFORM

SARBU, A.

Abstract: *The usage of telemedicine applications can generate terabytes of CDRs (Call Detail Records) everyday, corresponding to the voice and data traffic exchanged. The paper focuses on proposing the optimal data architecture and data modelling solution for an analytic platform designed to process CDRs generated by telemedicine applications. An analysis of CDR data could enable valuable insights that may represent a future basis for key decisions to improve the medical services offering.*

Key words: *Data warehouse, OLAP Cube, Call detail records, Data mining, Data mart*

Authors´ data: PhD Student **Sarbu**, A(nca), Polytechnic University of Bucharest, Splaiul Independentei no 313 sector 6, Bucharest, Romania, sarbuanca@gmail.com

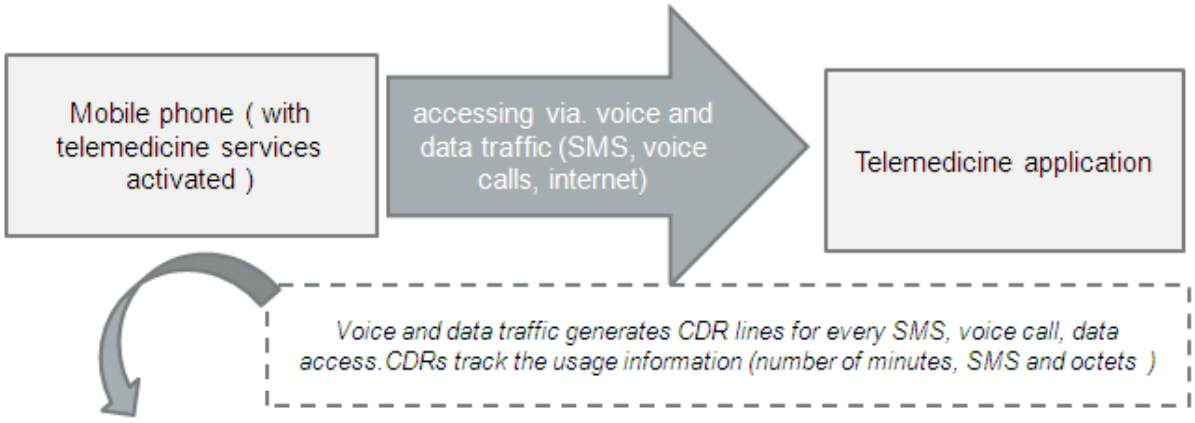
This Publication has to be referred as: Sarbu, A(nca) (2013) Optimal Data Architecture for an Telemedicine Analytic Platform , Chapter 37 in DAAAM International Scientific Book 2013, pp. 647-654, B. Katalinic & Z. Tekic (Eds.), Published by DAAAM International, ISBN 978-3-901509-94-0, ISSN 1726-9687, Vienna, Austria
DOI: 10.2507/daaam.scibook.2013.37

1. Introduction

Telemedicine is the domain that delivers interactive health care at a distance by using various modern telecommunication technologies. Due to the fast progress and growth in this domain, telemedicine services came to cover a large part of the medical areas (from general health delivery to specialist care delivery) and overcome the distance barriers in order to deliver a multitude of medical services to less accessible geographical areas.

An example of a simple telemedicine service is video consultation thru videoconferencing. Medical information is exchanged from one sight to another via electronic communication and the means of providing these services could be mobile applications. .If the telemedicine services are activated on the mobile phone, a user can schedule a videoconference in order to be consulted by a health care professional. The interactions between the patient and the health care professional intermediated by the telemedicine services generate voice traffic (e.g. if after the consultation a recipe will be issued, an SMS with the recipe will sent) and data traffic (e.g. if the personal medical record is accessed, internet traffic is generated) (Graschew & Rakowsky, 2011).

Every interaction (such as an SMS, call, internet access etc.) can be traced, as it will generate a unique CDR line, see Fig. 1.



CDR_DateStart	CDR_SupportTrafficId	CDR_IMSI	CDR_NoUnits	CDR_IMEI	CDR_Geoloc1	CDR_Geoloc2
15/08/2013 00:13:22	17	208010300112098	27	1216100942030	25856	31476
15/08/2013 11:08:30	17	208010300112099	680	1216100942030	38659	23644
15/08/2013 13:52:51	34	208010300112096	23	1216100942030	38700	24656
15/08/2013 13:52:52	34	208010300112096	33	1216100942030	38700	24656
15/08/2013 15:04:09	34	208010300112096	1	1216100942030	38700	24656
15/08/2013 17:11:54	34	208010300112097	1	1216100942030	38659	23656
15/08/2013 18:59:01	34	208010300112097	34	1216100942030	38659	23656

Fig. 1. Interactions with the telemedicine application, generating CDRs

Telemedicine researches nowadays focus more on the complexity of the services that can be delivered and on the development of new paradigms of healthcare delivery (Forbes & While, 2009). One area that has been neglected, but might

provide useful insights, consists in the analysis of the data generated by a telemedicine application. Determining which medical service is most used in a certain geographical area, over a particular period of time, could help in linking the frequency of accessing a specific medical service with some environmental factors. This is only a simple type of analysis, but furthermore spatial data mining algorithms, such as discover-spatial-characterization algorithm, can be used for example to determine the level of demand for specialized medical services or other interesting properties of the geographical areas. In this case a non-spatial attribute such as the use of standard medical services access rate could be chosen as a relevant indicator.

In order to enable such an analysis, the first step is to identify the data sources needed (CDR sources) and then apply the data extract-transformation-loading process that will ultimately feed the analytic DBMS. The paper focuses on presenting the optimal data architecture for the final analytic DBMS, thus preparing a platform for the future model building and spatial, statistical and predictive analysis.

2. Data Modelling

Telecommunication companies store call information such as source and destination phone numbers, type of traffic exchange (voice/data), type of telephonic service used (voice call, SMS, MMS, videoconferencing etc.), call duration etc. This information is known in the telecommunication industry as call detail records (CDRs). Also they dispose of information regarding active antennas and their coordinates in order to geographically determine the location of the phone number who has originated/received a call (Ćamilović et al, 2009). With the purpose of defining the best suited data architecture for creating a platform designed to analyse usage data generated by telemedicine application, the first step is to create a Staging database that will contain CDR and antennas data sources. Depending on the complexity of the analysis more data sources can be added such as a billing data source.In this way usage information can be correlated with the amount billed for the services a user has accessed. The Staging database will act as an intermediate storage area that will further feed the Data mart database. The data from the Staging database is raw data, with no aggregation as it can be seen in Fig 2.

The data that should be included in the Staging area corresponds to: Voice CDRs, Data CDRs and antenna localization information. Also specific type of services should be identified and marked in the incipient phase of the design. For example, if the application contains a simple service such as the possibility to verify the detailed medical history upon request, this service should be uniquely identified by a code and marked in the database. If a user will request this service an SMS with the patient medical history could be generated by the application and sent to the user. The staging information can suffer further modifications in order to be mapped to a specific telemedicine application and the services it provides. If for example there is an already defined pool on MSISDN (Mobile Subscriber ISDN Number) that generates the input/traffic of the application, then it should be also stored in the Staging area and afterwards used to create flags.

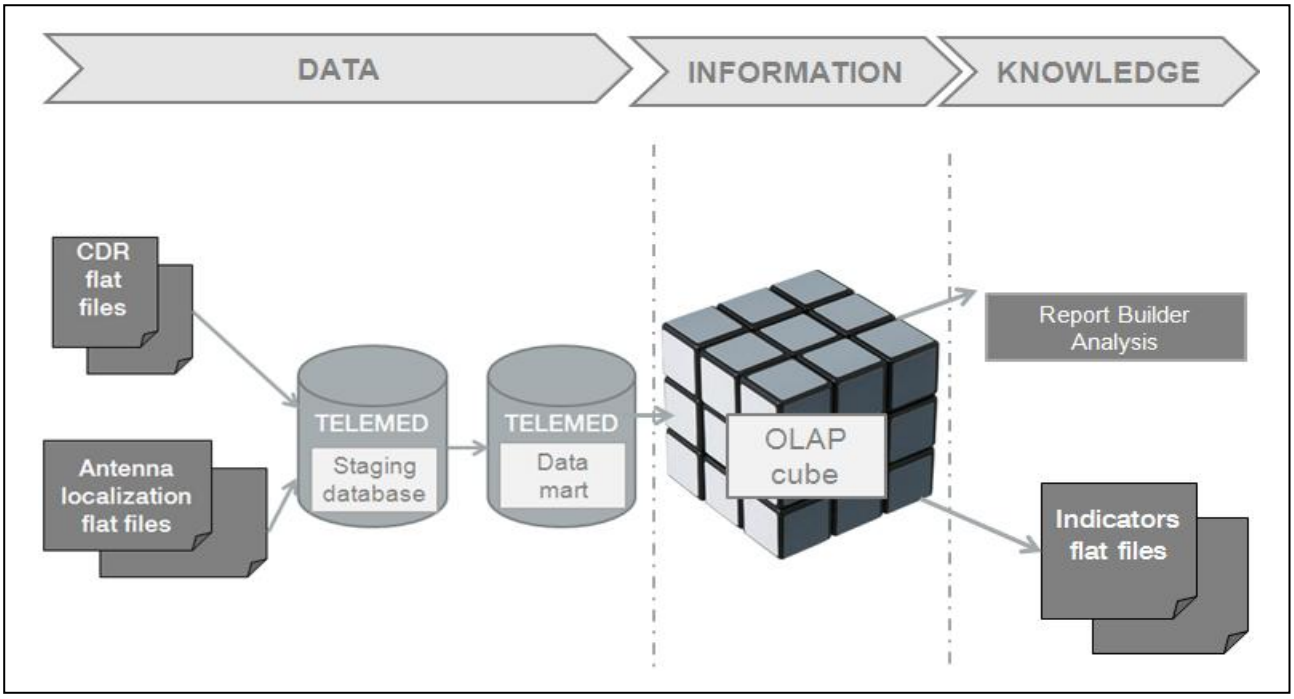


Fig. 2. TELEMED Conceptual schema

Field	Field description
CDR_TypeID	Identifier of the CDR type. Could correspond to incoming, outgoing, redirected.
CDR_DateStart	Date of the beginning of the call.
CDR_SupportTrafficId	Identifier of the type of support traffic.
CDR_IMSI	International Mobile Subscriber Identity or IMSI is a unique identification associated with all cellular networks. An IMSI is usually presented as a 15 digit long number. The first 3 digits are the mobile country code (MCC), which are followed by the mobile network code (MNC), either 2 digits (European standard) or 3 digits (North American standard). The length of the MNC depends on the value of the MCC.(1) The remaining digits are the mobile subscription identification number (MSIN) within the network's customer base.
CDR_MSISDN_Generating	MSISDN is a number uniquely identifying a subscription in a GSM or a UMTS mobile network. Represents the MSISDN of the subscriber generating the call.
CDR_MSISDN_Corresponding	Represents the MSISDN of the subscriber receiving the call
CDR_NoUnits	Number of generated units (seconds, octets, SMS etc.)
CDR_UnitsType	Type of generated units(seconds, character etc.)
CDR_IMEI	The International Mobile Station Equipment Identity or IMEI is a number, usually unique, to identify 3GPP (i.e., GSM, UMTS and LTE) and iDEN mobile phones, as well as some satellite phones. The IMEI number is used by a GSM network to identify uniquely valid devices.
CDR_SoftVer	The software version of the application corresponding to a IMEI.
CDR_TermReason	Code corresponding to the cause of ending the call
CDR_CentrerInvoice	Invoice center necessary to identify if there is any roaming.
CDR_Geoloc1	Code corresponding to a geographical area.
CDR_Geoloc2	Identification of the coverage area for a network.

Tab. 1. Example of Voice traffic information that could be integrated

In order to facilitate the geo-spatial analyses that are in the scope of the application, there is a need to integrate antenna coordinates, aside from CRD data. The geocoding is World Geodetic System 1984 (WGS 84) and the Lambert 2 projection.

The data contains antenna localization information such as country, city, street, area etc. By using the Lambert II projection the latitude and longitude coordinates can be translated into abscise and ordinate that will further help for the distances calculus. Based on this information it could also be determined how many antennas are active at SIM card level. Analysing the distance between antennas could be an important indicator to study the behaviour of a telemedicine application user.

The integration of the data could be done via Microsoft SSIS (SQL Server Integration Services) packages. In terms of data warehouse, the option was to use a data mart, as the focus is only on a subset of data (voice and data traffic analysis taking in consideration the geographical aspect) and not on all available data. As a result of the tree-stage process for designing a database (conceptual, logical and physical data models) a data mart was also proposed in (Ćamilović et al, 2009), as being the best suited data structure to store CDR information and the further use OLAP as a natural extension for obtaining the needed insights.

Field	Field description
GPRS_TypeID	Identifier of the CDR type.
GPRS_DateStart	Date of the beginning of the call.
GPRS_SupportTrafficId	Identifier of the type of support traffic.
GPRS_IMSI	International Mobile Subscriber Identity or IMSI is a unique identification associated with all cellular networks. An IMSI is usually presented as a 15 digit longnumber. The first 3 digits are the mobile country code (MCC), which are followed by the mobile network code (MNC), either 2 digits (European standard) or 3 digits (North American standard). The length of the MNC depends on the value of the MCC.(1) The remaining digits are the mobile subscription identification number (MSIN) within the network's customer base.
GPRS_MSISDN	MSISDN is a number uniquely identifying a subscription in a GSM or a UMTS mobile network. Represents the MSISDN of the subscriber generating the data transfer.
GPRS_AcessPointId	WLAN access point
GPRS_NoUnits	Number of generated units . The type of units could be seconds or octets.
GPRS_UnitsType	Type of generated units(seconds, character)
GPRS_IMEI	The International Mobile Station Equipment Identity or IMEI is a number, usually unique, to identify 3GPP (i.e., GSM, UMTS and LTE) and iDEN mobile phones, as well as some satellite phones. The IMEI number is used by a GSM network to identify uniquely valid devices.
GPRS_SoftVer	The software version of the application corresponding to a IMEI.
GPRS_VolIN	The total volume of in-coming data in octets
GPRS_VolOUT	The total volume of out-going data in octets
GPRS_TermReason	Code corresponding to the cause of ending the call
GPRS_CentrerInvoice	Invoice center necessary to identify if there is any roaming.
GPRS_Geoloc1	Code corresponding to a geographical area.
GPRS_Geoloc2	Identification of the coverage area for a network.

Tab. 2. Example of Data traffic information that could be integrated

In the dimensional data model proposed there is a fact table (central to the data warehouse) named tblCall, as it is presented in Fig.3. The choice of creating this table was because it contains multiple dimensions values (measures) such as number of calls, number of seconds, number of octets etc. In this case, the fact table will not have only one level of dimensions tables so this is why snowflake schema is better suited than the star schema for the modelling (Adamson,2010).

Because the information concerning the equipment and call type will not be used frequently in queries, separate tables are created: tblEquipement and tblCallType. As an advantage, this removes the duplication that occurs in case of duplicated information about equipment used to make all the calls.

However, as downsize the queries will be slower because they will require an extra join in this case (snowflake schema contains normalized tables opposed to star schema where all tables are de-normalized).

Although star schema is the simplest data warehouse schema (contains only a single dimension table for each dimension) and the query performance is better, as the information about the medical partner that provides the medical services is very sparse (there is no much data about it in many cases) it is a good choice to create another table for the dimension provider. Also this is a choice that better supports maintenance and change. So it is easier to modify it in case the analysis will have a new direction or will require new data. (Adamson,2010).

As the amount of data that needs to be stored has a big volume, the choice of implementation is to use Microsoft SQL Server 2008 R2 with Analysis Services. This technology also supports choosing as a logical schema the snowflake schema, as it is the most efficient way to populate the Analysis Services cube.

As this implies a large scale implementation, surrogate keys are used to ensure the data warehouse expansion over time. Instead of using business keys (antenna_id, imsi, end_user, service_id etc.), surrogate keys are considered as primary keys for all the dimensions. Business keys are attributes and surrogate keys are typically integer and are maintained by the staging process.

The advantage of using surrogated keys is ensuring the maintenance of the data integrity. This supports the resolution of namespace collisions in case it will be wished to combine multiple sources of data. It is also useful to have surrogate keys in case some business keys will be reused (Celko,.2010).

However, there is one downsize to using surrogate keys: the complexity and time costs related to loading the fact table (tblCall).

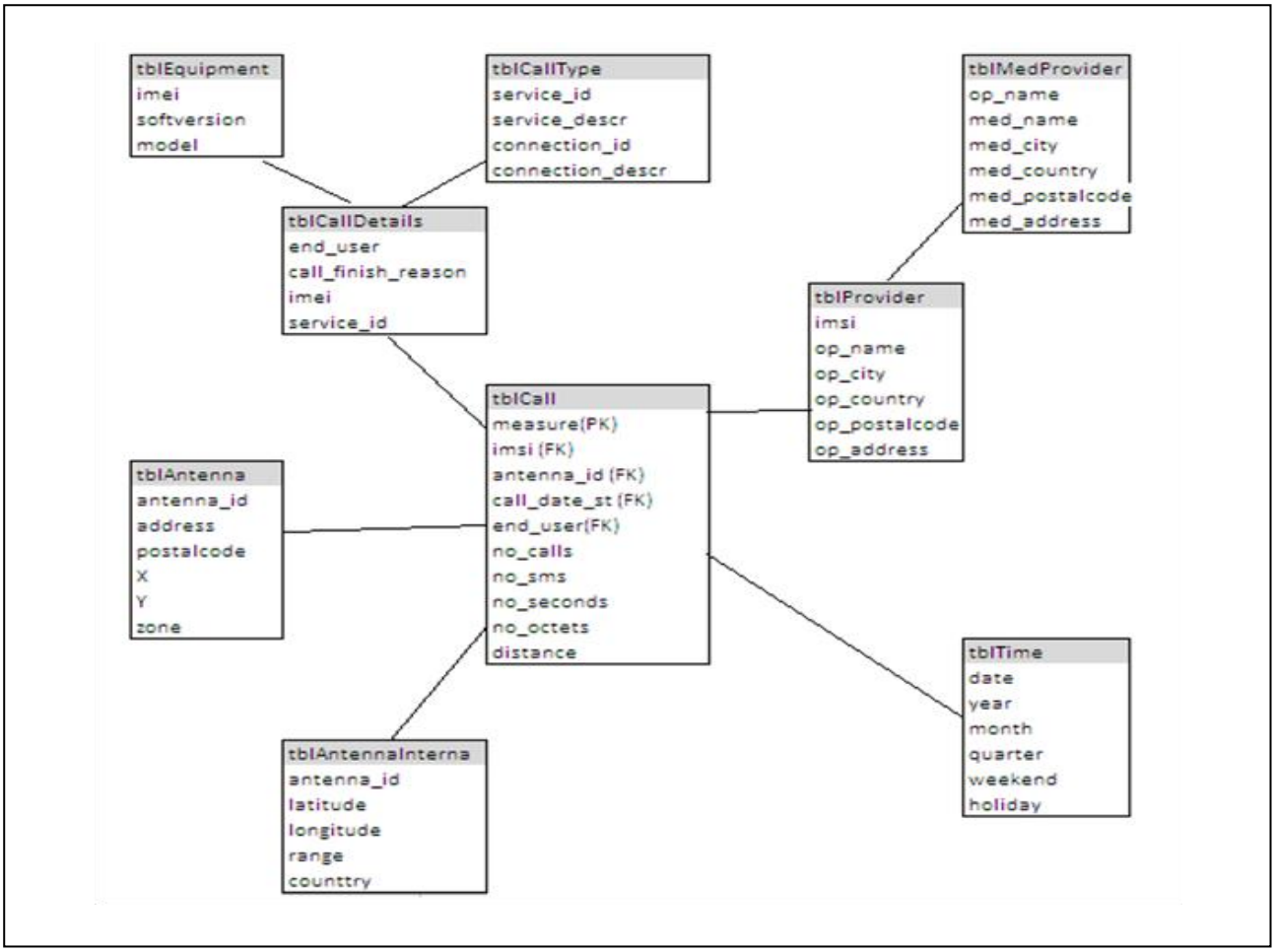


Fig. 3. TELEMED Data mart snowflake schema

3. Type of OLAP Analysis

The dimensional model will aliment the OLAP cube and allow several areas of analysis.

A first study could be made on the intensity with which a certain telemedicine service, for example teledermatology, has been used in a diversity of geographical area. This type of analysis could be useful to identify if there would be a need for a specialized point of service on a site with a high demand, over a constant period of time. If an analysis of evolution over time of the intensity usage is included, it is possible to identify if we are dealing with an isolated case. In this situation, further analysis and investigation could be done for the search of root causes. This aggregated historical data can be stored in one table. This will enable periodical comparisons to be done for example, on a monthly basis.

In order to determine the intensity, in case a voice traffic exchange, the number of calls should be analysed, together with the duration of the call (seconds).In case of a data traffic , the number of requests and volume of in-coming and out-going data(in octets) should be studied.

Another analysis could be made on the correlation between users that have accessed the services between a specific action radius (a determined geographical area) and users that have accessed the services always from a fix point.

4. Conclusion

As stated in (Islam R. et al, 2009) future research trends and challenges focus on interoperability, medical sensors, medical robots, human-machine interface,Micro-Electro-Mechanical Systems, cellular technologies etc. Machine-to-Machine(M2M), technologies that allow both wireless and wired systems to communicate with other devices, also gained a big importance as many of the telemedicine services are M2M based. Most of the research efforts nowadays focus on developing more services and on improving the solutions for the current ones, by using better performing technologies, optimized architectures etc.

This paper focuses on a line of research that has not been yet pursued, but that could have a big impact over the optimization of telemedicine services offering.

As telemedicine domain is vastly evolving, a solution that analyses the big volume of generated data is a must. The paper answers the first challenge of developing such an analytic platform: designing and modelling the optimal data structures that are ready to respond to any investigations regarding the usage and geolocalisation of telemedicine services.

Based on studies about traffic performed for a multitude of geographical areas, important insights can be discovered. Responses to questions such: where a certain service of telemedicine is more used? during the last 2 months where was the most demand of telemedicine services? are telemedicine services used intensively on roaming? , can help in the decision making process for improving medical services offering.

5. Acknowledgements

Daniela-Anca Sarbu thanks prof. univ. dr. ing. Mircea Stelian Petrescu for the continuous support and guidance in the study area of database structures and database modelisation.

6. References

Adamson, C.(2010), *Star Schema The Complete Reference*, McGraw-Hill Osborne, ISBN-10: 0071744320

Ćamilović, D.;Bečejski-Vujaklija D.;Gospić N. (2009)., *A Call Detail Records Data Mart: Data Modelling and OLAP Analysis*, ComSIS Vol. 6, No. 2, December 2009, ISSN: 1820-0214

Celko, J., (2010) , *SQL for Smarties: Advanced SQL Programming, 4th Edition*, Morgan Kaufmann, ISBN: 9780123820228

Forbes, A.;While, A. (2009). *The nursing contribution to chronic disease management: a discussion paper.* International Journal of Nursing Studies. 46, 1, (January 2009)

Graschew, G.; Rakowsky, S. (2011). *Telemedicine Techniques and Applications*, InTech, ISBN: 978-953-307-354-5, Janeza Trdine 9, 51000 Rijeka, Croatia

Islam R.; Begum R.; Ali S. (2009), *Handbook of Research on Modern Systems Analysis and Design Technologies and Applications*,pp.584-608, IGI Global ,ISBN:9781599048871