The double wave data warehouse lifecycle model

D. P. du Plessis & T. McDonald
Department of Computer Science and Informatics, University of the Free State, South Africa

Abstract

The data warehouse has played a very important role in the transformation of the state owned telecommunications company in South Africa. This transformation process concentrated firstly on targets set by government during the exclusivity period and secondly, on the preparation for competition. A new Double Wave Data Warehouse lifecycle model was developed and followed to fulfil the immediate need for information through the building of a DW and BI solution in the company. It consisted of two iterations for the development of a DW and BI solution. Wave one concentrated on the rapid implementation of a Business Critical Information Solution. In this regard management information on bad debts and debtors days were considered as Business Critical Information Solution needs. Wave 2 concentrated on modelling the ongoing requirement into a permanent dimensional model.
Keywords: lifecycle model, data warehousing, telecommunications BI, business strategy, business intelligence strategy, developing countries.

1 Introduction

South Africa as a developing country has, since 1994, gone through a process of privatisation of some of the state departments. The Department of Post and Telecommunications was one of those departments. In order to obtain exclusive rights, for a limited period, as the sole landline provider, government has set several targets for the new telecommunications company.

The exclusivity period was to last for five years, until May 2002, but could be extended to six years, if the telecommunication operator met network rollout and service targets. The rollout targets included doubling its subscriber access lines

WIT Transactions on Information and Communication Technologies, Vol 38, © 2007 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA070111

by 2.7million; installing 120,000 new public telephones; connecting 3,200 villages for the first time and providing service to more than 20,000 priority customers such as schools and clinics. The exclusivity period was intended to allow the company to expand the network as rapidly as possible in order to facilitate universal access and to move towards universal service. The agreement left the telecommunications provider with the challenge to plan and manage the implementation targets set by government, while at the same time preparing for competition which would start at the end of the exclusivity period.

2 Background

A data mart was developed to manage the spare infrastructure in order to sell telephone services in areas where infrastructure is available. This data mart helped the sales team to create a solution which provided addresses of possible new customers in the under serviced areas where there was telecommunication infrastructure. This new solution therefore reduced the orders which needed to wait for new infrastructure to be built. The sales channel was unfortunately not the only channel for new orders. There were also customers who fell into the higher income group who went to the front office to apply directly for a telephone service. Some of these customers were existing customers who moved from one city or town to another. These customers could afford the telephone service and had to be accommodated, because they are maintaining or improving the cash flow in the company. It was important to make infrastructure available to these customers, because of the company's target to double its access line.

The company sometimes needed to keep existing customers although they not the best paying customers. A solution was needed to point out the poor paying customers in areas where there were customers with a good credit record waiting for infrastructure. Although the replacement of a poor paying customer with the new customer did not increase the number of lines, it helped to maintain or improve the cash flow of the company. Replacing customers who could not afford the service and stay in the under serviced area, with customers that can afford the service and stay in an already serviced area could not be maintained, because one of the targets was to install services in under serviced areas. The unemployment rate in the under serviced areas of South Africa is high and made it very difficult for these people to afford a telephone service. To accommodate these people the company had to relax its credit vetting policy. Relaxing the credit vetting policy had a negative influence on debtors which could lead to an increase in bad debts.

Bad debts are written-off from time to time and then given to lawyers to collect. The bad debts which they collect are written back as "bad debts recovered". Targets were set to measure the percentage bad debts recovered by the attorneys which was also a requirement for the data warehouse. Bad debts are risky to any company and is therefore one of the indicators used by investors to evaluate their risk to invest in the company.

Before the landline prepaid service was implemented in South Africa, all landline services were post-paid. This meant that a customer only received a bill

for the service rendered and calls made at the end of the month. The bill normally had to be paid by the seventh of the next month. Debtor's days were calculated using the billing date. The payment process worked as follows: before a customer's service is installed, a deposit was due of R250 plus installation fee. If a customer's line was blocked for international calls and R3000 plus installation fee if not. The customer's service was suspended for outgoing calls if the outstanding balance for longer than thirty days reached the value of the deposit paid. If it exceeded sixty days the service was suspended for incoming calls too. When the outstanding balance was higher than the deposit for ninety days, the service was dismantled.

Collections were done using the debtor's management BI solution as explained below. Collection agents were receiving incentives for collecting debts before it is written-off as bad debts. The process for collections worked as follows:

- the collection agent receive the debtors management report;
- the agent lift the suspension of the debtor temporarily;
- the agent makes a call to the debtor and negotiate payment dates;
- payment dates are inserted on billing system;
- payment dates are extracted into the data warehouse for collection reporting;
- the agent uses the collection reporting to follow-up on promised pay dates by a reminder a day before the agreed payment date.

Incentives were only paid to agents when debtors kept their promises. This was to avoid the payment of incentives when the payment was not due to the agent's collection effort. Because of the small percentage of debtors who paid on the exact day of agreement an arrangement was made that when the payment was made five days after the agreed date, the agent still get the incentive, otherwise the agent had to start with the collection process all over again.

3 Data warehouse implementation using the double wave data warehouse lifecycle model

The Double Wave Data Warehouse (DWDW) lifecycle model (see figure 1) was used to develop the BI and DW solution which was used to manage the debtors based on the above mentioned four requirements.

Each step of wave 1 of the model will now be explained and will be followed by the steps of wave 2.

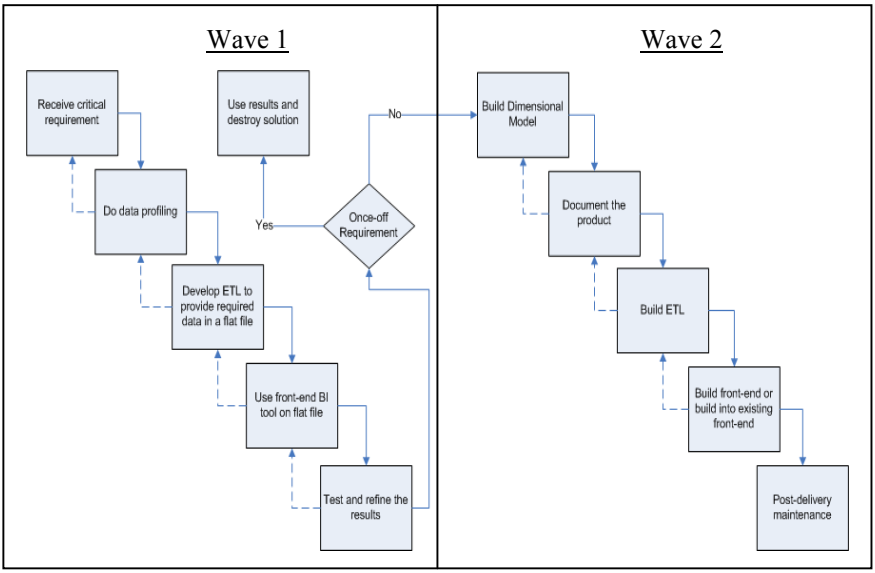

Figure 1: The double wave data warehouse lifecycle model.

3.1 Receive critical requirement

A critical requirement was received from the credit management division in the company. Failing to implement this solution could mean huge cash flow problems for the company. A solution was needed urgently and could not wait for months or years for a data warehouse to be built. The first wave of the DWDW lifecycle model was needed for this Business Critical Information Solution (BCIS).

For example, a requirement was to have a BI solution where debtors could be categorized according to the age of their oldest bill. There should be a category for thirty days, sixty days, ninety days and more than ninety days. Debtors were allocated to these categories based on their oldest bill. It was also important to know when a debtor was in a more than ninety day category, what was the amount of debt ninety, sixty and thirty days. The credit management department concentrated on the "more than ninety days" debtors to do the collection of those debts first. However, it sometimes happened that the debtor paid the older debts, but the current debts were growing rapidly. In some instances the debtor's telephone service was not suspended as the policy required. A debtor could also have more than one telephone service and the debtor sometimes paid for the one service and not for the other one. It was therefore important to report debtors on an account number and not on the telephone number.

While collecting the requirements from businesses it was realized that it was easier for business people to explain their requirement by means of a report or spreadsheet. Because business people are thinking in terms of a report, it was decided to leave business to explain them in the way they were comfortable with.

That has helped the DW team to design the front-end, while collecting the requirements. One of the biggest risks for a software project to fail is the fact that it is hard for a customer to relate to how the final solution would look and work.

When building a house it is easy for the customer to relate because the architect can show the plan of the house to the customer. The plan of the house looks exactly like the end product, just a smaller version. The data warehouse team therefore decided to design and built the front-end as represented in figure 2 below. At this stage the source data was not available. The data warehouse team therefore used test data to complete the design and built of the frond-end. The file hosting the test data was kept to serve as a template for the flat file with the real data.

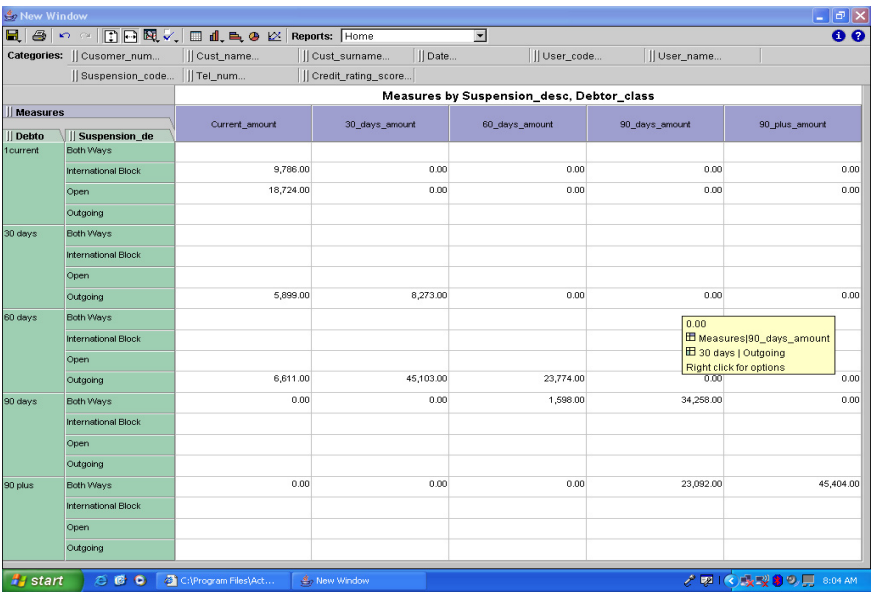

Figure 2: Data Cube for debtors management solution.

The rows in Figure 1 represented the different debtor's categories as well as the suspension descriptions, while the columns indicated the aging of the outstanding amounts. There were other dimensions like customer number, customer name, telephone number, etc, available in the cube which is not currently shown in figure 2, but are available to the customer using the life cube.

3.2 Do data profiling

Wikipedia encyclopedia [3] defines data profiling as a process whereby one examines the data available in a source database and collects statistics and information about that data. The purpose of these statistics is as follows:

- find out whether existing data can easily be used for other purposes;
- provide metrics on data quality, including whether the data conforms to company standards;
- assess the risks involved in integrating data for new applications, including the challenges of joins;
- track data quality;
- assess whether metadata accurately describes the actual value in the source database;
- understanding data challenges early in any data warehouse project, so that late project surprises are avoided. Finding data problems late in the project can incur time delays and course project cost to overrun.

This process was therefore important to ensure that the Extract Transform and Load (ETL) were done with the right and best data. The billing system in the company was up to date when it came to billing information, but a better source was needed for information on the suspension of telephone numbers. Data profiling has revealed that the exchange database was the best source for suspension information and that the telephone number was the only common field by which the two sets of information could be combined.

3.3 Develop extract, transform and load process to provide required data in a flat file

The first part of an ETL process is to extract the data from the source systems. Most data warehousing projects consolidate data from different source systems. Each separate system may also use a different data organization format. Common data source formats are relational databases and flat files, but may include non-relational database structures such as IMS or other data structures such as VSAM or ISAM. Extraction converts the data into a format for transformation processing.

The load phase loads the data into the data warehouse. Depending on the requirements of the organization, this process ranges widely. Some data warehouses overwrite old information with new data and these dimension tables are called type 1 slowly changing dimension tables (Kimball [2]). More complex systems can maintain a history and audit trail of all changes to the data, the type 2 and 3 slowly changing dimension tables are used for these purposes (Kimball [2]).

The ETL process needed to put the data in a flat file for the first wave of the Lifecycle Model. The data was not only coming from different tables in the billing system, but also from the exchange database. The suspension type, and user code and name of user that has changed the status, came from the exchange database. All data sets were combined to create a flat file to report from.

3.4 Use BI front-end tool on flat file

BI front-end tools like Actuate, Business Objects, Essbase, etc. can use a flat file as a data source. The main aim of the first wave of the DWDW lifecycle model was to fulfil the immediate information need and not to optimize the solution. The frond-end was already designed, while the business requirements were received form the business customer. The main aim of this phase was to link the front-end with the source data, which is at this stage in a flat file. The lifecycle up to this point has taken approximately ten days. This meant that there was a solution in a very short period of time. The solution was ready for acceptance testing.

3.5 Test and refine the results

Testing the interim solution required a lot of data profiling. At this stage of the DWDW lifecycle model the sources of the necessary data for the flat file could still change. Version control of the ETL jobs is very important. There should at least be two versions of the ETL at any given time during the development of the solution. Most ETL tools make provision for development comments. Development comments were on the ETL job to make future corrective, adaptive and perfective maintenance easier.

3.6 Is requirement once-off?

Not all the information needed was an ongoing requirement. If it was a once-off requirement, the flat file was destroyed, but the ETL jobs were stored and could be reused for new developments, or the flat file could be recreated if required in future. If a permanent solution was required, the project team started wave 2 of the DWDW lifecycle model.

3.7 Build a dimensional model

Figure 2 represents the data model for the debtor management solution. The dimensional model in data warehousing consists primarily of two kinds of tables, namely fact and dimension tables. Kimball [2] defined a fact table as the primary table in a dimensional model where the numerical performance measurements of the business are stored. This is illustrated in figure 3 in the Debtors_fact and Collections_fact below. Debtors_fact stores the outstanding amounts for debtors splitted up in the four different outstanding time frames. The Collections_fact stores the amounts collected. Both of these fact table have foreign keys that link to dimension tables. Kimball [2] defined a dimension table as a integral companion to a fact table. The dimension table contains the textual descriptors of the business. Dimension tables can have many columns or attributes. These attributes describe the rows in the dimension table. Dimension attributes serve as the primary source of query constrains, groupings and report labels.

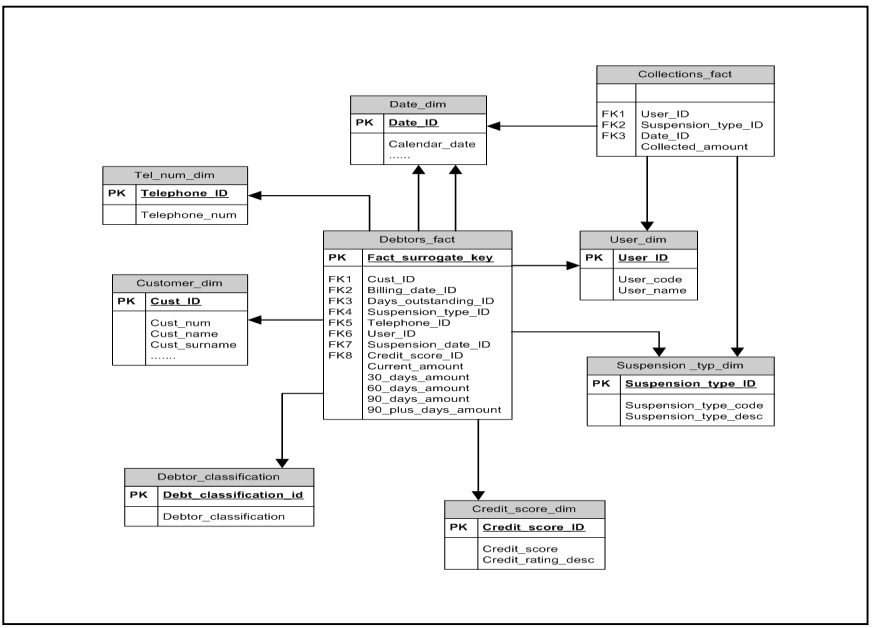

Figure 3: Data model for the debtors management solution.

3.8 Document the product

Post delivery maintenance of any software product needs documentation. For data warehousing, design and end-user documentation are needed. Wikipedia encyclopedia [3] describes this documentation as follows:

Design documents tend to take a broad view. Rather than describing how things are used, this type of documentation focuses more on the why. For example, in a design document, a designer would explain the rationale behind organizing a data structure in a particular way, or would list the member functions of a particular object and how to add new ones to the code. It explains the reasons why a given class is constructed in a particular way, points out patterns, and even goes so far as to outline ideas for ways it could be done better, or plans for how to improve it later on.

The **user documents** describe each feature of the solution, and the various steps required to invoke it. A good user document can also go so far as to provide thorough troubleshooting assistance. It is very important for user documents not to be confusing, and for them to be up to date. User documents need not be organized in any particular way, but it is very important for them to have a thorough index and a constant format in a particular company/organisation. Consistency and simplicity are also very valuable. User documentation is considered to constitute a contract specifying what the BI solution will do.

3.9 Build ETL

The ETL in the second wave of the DWDW lifecycle model concentrates more on the transform and load of the ETL process. The first wave takes care of the extraction and transformation which concentrate on changing data types from different source systems into the same data type and to make the linking of his data sets, coming from different sources, easier for loading into the flat file. In most of the cases the wave one flat file was used for staging for transformation in wave two, but this time the transformation goes about transforming data to load into the dimensional model. When loading a dimension table in a dimensional database, it is easy to do a distinct select SQL statement and load it directly into the dimension table, while the ETL tool creates a new surrogate key for every new row loaded into the table. With loading the fact table, surrogate keys from the dimension tables need to be loaded into the fact table as foreign keys. That means several lookups need to be done for the transformation from a flat file to a dimensional database.

3.10 Build new front-end or build into existing front-end

This phase is required when the new front-end is built in different software than the front-end used in the first wave. This phase was used because the business people were not familiar with the new BI front-end tools. The DW team therefore used MS Excel as the front-end tool, because that was the tool that the business was familiar with. Excel was systematically replaced with the new BI tool as the BI maturity of the company improved Du Plessis and McDonald [1].

3.11 Post delivery maintenance

A good data warehouse is a data warehouse that is continuously expanding. The type of enhancements to a data warehouse consists of adding new subject areas (adding information about new business areas), adding new analytical capabilities, merging data warehouses and adding new sources of information. All these require additional resources and the ability for effective integration. All these generate additional maintenance activities.

Post delivery maintenance in a data warehouse environment can therefore become a big challenge. In the event of adaptive and corrective maintenance it might require that the tables are reloaded. When a new dimension is added it is in some instances unavoidable to reload the fact table. Naming conventions plays an important role in the maintenance of a data warehouse. This gives the maintenance team an indication of the business area to which the fact or dimension table belongs to, as well as the conformed dimensions and fact tables that are shared amongst the different business areas.

4 Conclusion

The DWDW lifecycle model was developed to simplify the process of implementing the data warehouse. In developing countries, other investment

needs take priority over the implementation of a data warehouse. It is therefore important for the data warehouse team to ensure continuous adding of value as the data warehouse is growing. The DWDW lifecycle model, through the use of wave one, ensures a quick response on BCIS. Wave two ensures the optimisation of the solution by modelling the solution into a dimensional model, which is the optimal design for a database, used for large queries and online analysis. Documentation in the last wave improves the maintainability of the solution.

References

[1] Du Plessis, D. & McDonald, T., Challenges in building and maturing of a telecommunications business intelligence solution in a developing country (To be published in IRMA 2007 proceedings), 2007
[2] Kimball. R., The Data Warehouse toolkit Second Edition: Developing and Deploying Data Warehouses. New York: John Wiley, 2002
[3] Wikipedia encyclopedia http://en.wikipedia.org/wiki/Data_profiling, 2006