

Data transformation for Business Intelligence: An Empirical Study

* To mask sensitive information, Dog Office and Dog Grant used in this report are fictitious.

Abstract

The Dog Office (DO) administers the Dog Grant (DG). There are six programmes under the DG. Data collected from the administration of these programmes resided in separate databases leading to inefficiencies in data analysis. This report documents the process of consolidating the data into a single database, leading to the ability to track, in real-time, the utilisation of the fund and significant reduction in time spent preparing management reports.

1. Introduction

The Dog Office (DO) administers the Dog Grant (DG). This is a sectoral fund to aid the dog industry in training and talent development. There are six programmes under the DG. Dog companies submit applications to the DG administration for funding support under one of the six available programmes.

As part of regular reporting to DO board and senior management on the usage of the DG, regular reports are prepared. These reports are in the form of monthly, quarterly and annual reports. The reports provide DO board and senior management with, among other things, the progress of utilisation of the DG and the beneficiaries of the DG. Depending on which programme is applicable, the beneficiary could be either a dog company or an individual employed in a dog company. Whenever necessary, DO Board decides if the fund should be topped up or if there is to be any strategic reallocation of the fund across the six programmes.

2. Problem Statement

Data collected from the administration of the programmes resided in separate databases. Each database recorded a set of information depending on which programme was being applied for. The different programmes under the DG require a slightly different set of information as part of the application process.

Preparation of management reports was a tedious and time-consuming process. Information was recorded inconsistently. For example, there was no standard way of recording date information. At times, it was recorded as Aug 2014 and other times, it was 16-08-2014. It was inefficient trawling through the data to glean insights. In addition, the information resided in silos; in separate databases. The DG administrator was unable to have an overview of

how a company or individual had benefitted from the DG.

To ensure consistency in the data, the data had to be changed to the same format. For example, since "P/L", "Pte. Ltd." or "Pte Ltd" were being used interchangeably, it had to be amended to a standardised form. Records in the individual databases were transferred into a single new database.

3. State of the Art

The evolution of Information Technology

The work that Information Technology (IT) personnel has evolved over time. Where it used to be programming-based, it is now focused more on the integration and customisation of IT solutions to clients (Niederman and Mandviwalla, 2004).

There are now greater levels of consultation between IT and business personnel when it comes to mining big data (Pierce, 2012). This sustained, iterative process between business and IT ensures that the delivered products and solutions are feasible to the client.

Customer Relationship Management

Zahay *et al* (2012) discusses the types of customer-based data and the quality of Customer Relationship Management (CRM) systems developed from such data. The work focusses on CRM systems for financial services firms. In managing customer and stakeholder relationships, transactional data is essential to the Recency, Frequency and Monetary (in short, RFM) value calculation.

Data Architecture

It is essential that the data architecture adopts a comprehensive approach to integrating data. This approach allows stakeholders to know the customer (McKendrick, 2014). The creation of customer knowledge competence allows organisations to manage their CRM programmes strategically (Campbell, 2003).

In a fast-moving environment, real-time data is crucial. Real-time data integration moves information expeditiously from the data sources to the decision makers (McKendrick, 2014).

Mullins (2012) gives practical ways of database design to ensure usability. One common database problem is overstuffing of columns. When a single column is used to store what should be in more than one column, this design flaw can possibly make the database unusable. It may work for one application, but once the business requirements change, the database very quickly becomes obsolete.

4. Solution: Description of Approach

Designing the database

Kimball's (2013) CRM case study was referred to in designing of the data warehouse. Across the six databases, one each for the six programmes, the common columns were identified. The column names were amended to reflect more accurately the details in that column. In designing the single database, business domain knowledge is required to understand the workflow of how these applications are handled. For details that were not applicable to the programme, the database does not require that information for the transaction. For example, the course/attachment end date is not applicable to programmes two, three and four. Another example is that Programme Four does not have an approved DG amount committed. This is due to the workflow process whereby all programmes except Programme Four first have an amount committed to that transaction and subsequently an amount that is disbursed. The committed amount and disbursed amount could differ.

Table 1 below shows the proposed columns of the single database. The greyed-out boxes indicate which columns are not applicable to the programme.

	Programme number					
Column headers	1	2	3	4	5	6
Reference number						
Trainer Name						
Course Name / Course of Study						
Course Type						
Max payable course/training fee						
Percentage of funding						
Course/Attachment end date						
Maximum grant per participant						
Name of DG participant						
Company (or previous Company) of DG participant						
Sector of Company of DG participant						
Place of attachment						
Total number DG participants						
Approved number DG eligible participants						
Approval date						
Approved DG amount committed						
DG amount disbursed						
DG disbursement date						
DG amount balanced						

Table 1: Proposed columns of the single database

Cleaning of data

Data cleaning was done using SAP Lumira software. Cleaning of data was necessary, as the data was inconsistent; both in value and in format. "P/L", "Pte. Ltd." and "Pte Ltd" were used interchangeably. For the single database, "Pte. Ltd." was used. Also, values such as those relating to company names were inconsistent; sometimes owing to spelling mistakes arising from past

human errors.

Transferring of data to single database

The data from each of the six programmes was transferred to the single database using SQL queries. The transfer was done in batches rather than at one single point in time. The batch transfer will be explained in the next sub-section.

Experimenting on the single database

Cleaning of data took up nearly 80% of the entire project's duration. During this time, the business continued to operate and there were management reports to prepare. To ensure the usability of the single database, the database was tested initially by transferring in batches the cleaned-up data, and using the single database to obtain the required information for reporting purposes. If the required information could not be obtained for reporting purposes, the design of the single database would have to be changed. Fortunately, due to good database design, the single database allowed the same analysis of data as what could be achieved from the six separate databases, albeit in a more efficient manner.

5. Discussion

Reduces time spent preparing management reports

The single database reduces significantly the time spent preparing management reports. Now that the data is integrated, the information comes from a single data source when previously there were six separate databases. The integration provides the Business Intelligence (BI) user a more robust way of analysing data. Previously, if a company receives funding under a programme, the BI user will not be immediately aware of funding under another programme. Only if the BI user goes to each of the other databases will the user be able to check if the company also receives additional funding.

Ensures data integrity

When the BI user keys in information to the single database, the input field will provide various suggested inputs for user selection. As the data had been cleaned, the suggested inputs are valid and will minimise human error encountered in manual typing. This also ensures data integrity.

Allows real-time monitoring

The design of the six separate databases was not done in accordance with industry's best practices. The databases did not contain directly the required information for management reporting. Whenever a management report is to be prepared, staff would prepare a preliminary document, different in format from the management report, so that the BI user can prepare the required management report. There are redundancies in the workflow process.

With the new database design, the workflow process is streamlined, as the BI user can conduct analysis of the DG utilisation in real-time. This allows DO Board to have up-to-date information of the utilisation of the fund. In an ever-

changing environment, it pays to have real-time data, so that business decisions can be made such as with regards to a top-up of the fund or strategic reallocation of the fund across the six programmes.

6. Conclusions and Further Work

This report documents the database transformation so that BI can be conducted more effectively on the DG. The literature review covered the evolving job scopes of IT personnel and how there is an increasing trend of greater interaction between Business and IT professionals. The author reviewed practical ways of database design, especially in the context of CRM.

From data previously residing in silos across various databases, a single database was built. The data cleaning and transferring process was documented. With a single database, the BI user could track, in real-time, the utilisation of the fund. In addition, there was a significant reduction in time spent preparing management reports.

With an understanding of the specific business domain, the process of designing a database serving the specific needs of the business can be applied to other industries. With a backend database in place, further work can be done to design a front-end application accessing the database. Research can be done on user interface and web design to ensure end-user satisfaction.

Bibliography

- [1] Niederman, F. and Mandvivala, M., *The Evolution of IT (Computer) Personnel Research: More Theory, More Understanding, More Questions*, Database for Advances in Information Systems, 2004.
- [2] Pierce, D., *Why Big Data is not a Death Sentence*, Database Trends and Applications, 2012.
- [3] Zahay et al, *Building the foundation for customer data quality in CRM systems for financial services firms*, Journal of Database Marketing & Customer Strategy Management, 2012.
- [4] McKendrick, J., *A New Era of Data Integration Dawns*, Database Trends and Applications, 2014.
- [5] Campbell, A.J., *Creating Customer Knowledge Competence: Managing customer relationship management programs strategically*, Industrial Marketing Management, 2003.
- [6] Mullins, C.S., *Avoiding Database Design Traps to Ensure Usability*, Database Trends and Applications, 2012.
- [7] Kimball, R. and Ross, M., *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modelling*, 2013.