



Data Warehousing Buy vs. Build

The Retail BIDA Solution
The Telco BIDA Solution

© Sean Kelly & Associates Ltd. 2009.

Publication Date: 23 June 2009

All rights reserved. No part of this document may be reproduced, stored in a retrieval system, or transmitted in any form of by any means whatsoever without the prior written permission of Sean Kelly & Associates Limited.

The information contained in this document is subject to change without notice. Although this information is believed to be accurate at the time of publication, Sean Kelly & Associates assumes no responsibility for the accuracy or completeness of this information or for this information being correct.

Sean Kelly & Associates makes no warranties, express or implied, relating to this document, or to any products or software described in this document.

Sean Kelly & Associates software referred to in this monograph may be used only under the terms of the Sean Kelly & Associates software license agreement.

Sean Kelly & Associates Limited

info@seankellyassociates.com

www.seankellyassociates.com

Sean Kelly & Associates (SKA) is a specialist boutique consultancy that specializes in the provision of data warehouse and CRM solutions and professional services to the telecommunications and retailing industries. Founded in 1992, and based in Ireland, the company employs 14 field consultants who each have, on average, 15 years experience of working in the business intelligence industry. Because the company has both business and technical expertise it has been possible to develop sophisticated solutions to real-world problems and to develop close and durable relations with leading companies in the telecommunications sector. The following list of clients includes only those where an SKA consultant provided the lead design capability to develop a full enterprise data warehouse. In addition to these clients there are many more in the telecommunications industry who have engaged the services of SKA to provide professional services on smaller-scale projects.

Contents

<i>Foreword</i>	3
<i>Advantages of Buy Option</i>	3
<i>BIDA Components</i>	3
<i>BIDA Deployment</i>	4
<i>BIDA Architecture</i>	4
<i>Productivity Guide</i>	4
<i>Methodology</i>	5
<i>Dimensional Design Features</i>	7
<i>Appendix A (Solution Scope)</i>	8

1. Foreword

The SKA Business Intelligence Data Architecture (BIDA) solutions for the telco and retail industries provides a fully functional data warehouse solution that can be delivered in a relatively short amount of time compared to the higher risk traditional approaches for custom-designed data warehouse architectures.

2. Background & Advantages of BIDA Solution

The BIDA models have been created to accelerate the development and implementation of a physical data warehouse schema in a telecommunications company. The models are based on the twenty years of experience that has been amassed at Sean Kelly & Associates Ltd in the design and deployment of telecommunications data warehouse systems around the world. The benefits of using BIDA are as follows:

- Use of BIDA *mitigates risk* when building a data warehouse, as most of the design issues that will be encountered have been anticipated in the solution.
- Use of BIDA *accelerates productivity* by provided a ready-to-use database schema that includes 4,300 data fields
- Use of BIDA incorporates SeETL, a *mapping tool* that stores the source to target mappings and generates the ETL specifications required to populate the model.
- Use of BIDA incorporates a range of *standard business reports* that are included as stored procedures.
- Use of BIDA *optimizes query performance* in the data warehouse by employing a two-tier dimensional design.

- Use of BIDA *facilitates the development of analytical applications* by incorporating aggregated and derived data tables that can be used to rapidly deploy complex applications such as customer segmentation analysis.
- Use of BIDA *facilitates the extension of the data warehouse* to include other vertical lines of business such as retail sales.
- Use of BIDA provides a *platform-independent design* that can be implemented on any database and subsequently migrated to a different database with minimal disruption.

3. BIDA Components

The BIDA solution provides 80% of the data structures that are needed by a telecommunications service provider in a data warehouse and optimizes these data structures for query and load performance. In addition, BIDA converts atomic level data structures in the data warehouse into business-aware derived data structures that can be used to create the analytical applications that are needed by the business and that justify the cost of investing in a data warehouse. The BIDA solution comprises the following components:

- **BI4All core data model** that is used as the basis for all of the vertical data warehouse design solutions provided by Sean Kelly & Associates Ltd. Both the core model and the telco vertical extension are contained in a dimension physical schema that includes type 1 and type 2 dimensions as well as fact tables and which conforms to second normal form (2NF).
- **BIDA vertical industry extension** that incorporates all of the telecommunications-specific data that will need to be populated on a telco data warehouse.
- **SeETL mapping tool** which is used to map the source to target data and to report on project progress. In addition this tool is used to generate ETL specifications for Informatica/DataStage or DataStage and can be

used to actually generate the ETL during the prototyping and testing stages.

- **BIDA Methodology and Techniques** that are provided to assist in resolving project management, design, productivity and performance issues.

- Review of project plan progress
- Authorization of changes to project plan (Change Control)
- Review of budget and resources
- Review of Issues Log
- Clarification of roles and responsibilities
- Critical design and scoping decisions

4. BIDA Deployment

The key advantages of the BIDA solution are to reduce risk and to increase productivity and the design intelligence that is embedded in the design will serve to minimize the risk involved and the resources required. Three key personnel will need to be made available by the client organization to implement the BIDA. These will include the following:

Project Manager: It is recommended that a Project Manager, with not less than 50% of their time allocated to this project, is assigned to supervise the progress of the project and to co-ordinate the interdependencies within the client environment and between the client the SKA design team and the Development team.

Data Mapper: It is recommended that a full time, or near full time, data mapper be made available who will be responsible for capturing and storing the mappings that define the data migration between the source systems and the BIDA data warehouse.

Data Investigator: It is recommended that a full time data investigator be assigned by the client to the project team who will be responsible for establishing the data definitions and data source details in respect of data that is to be populated on to the BIDA data warehouse. This task will require interaction with the many individuals who have specialist knowledge of the Trio system, the ODS system, the MDM product catalog and the business users who have specialist knowledge of data definitions and business rules. This analyst will also be responsible for the gathering of reporting requirements for this phase of the data warehouse.

In addition, it is proposed that the client will constitute a **Steering Committee** to be chaired by the Project Sponsor that will meet not less than once a month and as often as is needed. The Steering Committee will include, in addition to the Project Sponsor, the client Project Manager and the SKA oversight consultant, and as many other participants as are deemed necessary. The Steering Committee will review the following monthly items:

5. BIDA Architecture

The architecture of the BIDA solutions is based on a dimensional design that is more flexible to maintain, easier to navigate and provides better query performance than the traditional highly normalized data warehouse schema. The effect of the architecture that is used is to eliminate the intermediate layer that exists in the conventional third normal form (3NF) schema with the following advantages:

- Provides a practical physical schema that employs views and stored procedures and does not require the data warehouse architect to grapple with conceptual and logical data models that have to be later converted into physical designs
- Provides users with direct access to all data and does not confine the ad hoc user to artificially constructed star schemas
- Provides an advanced level of query performance on conventional database platforms (e.g. Oracle) that are architected for transaction processing
- Eliminates the need for a second data transformation layer where normalized data is de-normalized for use by end users

6. Productivity Guide

The BI4Telco solution comprises a total of 50 fact tables and over 100 dimension tables and the mapping tool provides a clear and transparent means of recording the mappings. In general, the time it takes to fully implement a BI4Telco data warehouse takes from 6 to 8 months with one SKA data architect and two internal resources at the client company to complete the design and testing of the schema. The ETL development work can commence in month three. The following guide may be used to estimate the effort

required. With regard to data mapping from source systems to BI4Telco the following table provides an indicative guide to the levels of productivity that have typically been achieved in similar projects.

Field Mapping Type	Mappings per day
Simple Migration	60-70
Data Translation	30-40
Data Calculation	5-10
Complex Semantic Transformation	1-5

7. Methodology

The following process describes the methodological approach adopted by SKA. Our approach will be to perform extensive prototyping. Our experience has been “If you cannot build the prototype you cannot build the real thing.” Therefore, we propose that build an extensive prototype from which to learn prior to building the real EDW and real ETL sub-systems.

SKA brings to the table the tools and techniques to build a large scale, and if the hardware is provided, full sized prototype prior to the development of a ‘production database’ or ‘production ETL subsystem’. Experience has shown that the development of the ‘production ETL subsystem’ and the cycle of change to the data model and change to reports and then change to the ETL subsystem is one area where many project have hit ‘stormy waters’. Hence the development of tools to avoid those stormy waters and to reduce the cost and risk of the development of the EDW by allowing the development of a large scale prototype prior to incurring the expense of the deployment of the production database and development of production ETL.

A high level summary of the tasks to be completed are as follows. We have extensive education materials that describe many of these steps in great detail which will be used to train and familiarize team members with the approach that we use.

7.1 Load substantial volumes of test data into the prototype staging area.

We will request substantial amounts of test data that is highly indicative of the production data to be loaded. We are aware there may be issues of security and privacy of data and that some contents of data might need to be modified to be used in the development environment. We propose that the data used in development be as close to ‘real’ as is possible given corporate policies that needs to be observed.

SKA will load this data into the prototype staging area using SeETL. The data in the staging area will be used as a valid sample of data from which to base the mapping process. It is required that valid samples of all data are placed into the staging area.

7.2. Mapping Source to Target.

We will document all mappings using the SeETL DesignTime Workbook. The SeETL DesignTime Workbook has evolved over many years and many projects to be a place to capture virtually all data that is required to define an ETL subsystem. It performs other tasks such as maintaining Business Objects Universes or printing Informatica/DataStage Programmer Specifications as well. Most data from the workbook can be loaded into any ODBC compliant database to be queryable. With SQL Server there are a series of reports in SQL Server Report Services available from the data loaded from the spreadsheet.

A major portion of the ‘Mapping’ exercise is the process of understanding the data that is in the staging area and determining where that data belongs in the target data model. The left hand side of the mapping spreadsheet is filled in by running a simple query against the Database catalogue to list all fields that are available in the staging area. The right hand side of the mapping spreadsheet starts out blank.

As the data in the staging area is understood and the proper location for that data in the target model is determined the right hand side of the mapping spreadsheet is filled in manually. There is space for ‘questions and comments’ in the workbook for follow up on fields that are not understood. All this data is available through the reporting provided with SeETL.

This process of understanding the data and filling in the right hand side of the spreadsheet continues for a period of weeks depending on how complex the mappings are and how many fields need to be mapped. The types of mappings that are particularly time

intensive are those where complex calculations must be defined and there is consensus required for those mappings to be completed. The generation of consensus can take many times longer than the mappings themselves. We will be fully aware of the issues of creating consensus when defining calculations. This process of mapping source to target in the SeETL DesignTime Workbook is typically the most critical task in the project. By making all mappings available through the reporting system can monitor the progress being made in detail.

7.3 Customization of the model.

The customization of the model is not started until portions of the mapping spreadsheet are determined to be 'solid'. That is, the Data Warehouse Architect (DWA) is confident enough that the target table will not change and therefore customization of the model can proceed with little likelihood of rework. The process is for the DWA to 'release' portions of the model to the DBA to perform the customization of the views over the base tables. During prototyping the views will be changed to present the 'target model' and the underlying tables will be changed to accommodate new fields. The DBA is expected to be a 'filter' to detect errors on the behalf of the DWA. As the mapping spreadsheet matures and more fields are mapped the target model can be customized in an overlapping fashion to create the final schema. In this way the model customization usually finished about one week behind the completion of the mapping spreadsheet.

7.4. Implementation of Prototype ETL.

The Mapping Spreadsheet also serves as the basis for implementing the prototype ETL. To implement the prototype ETL extra information needs to be defined such as the real key to be identified in the source data. The Prototype ETL will be implemented to populate the prototype EDW.

7.5 Analysis of the prototype DWH.

By populating the prototype DWH it will then be possible to determine any 'gaps' in the ability of the prototype EDW to support the reporting requirements. The developers of reports will be able to query the prototype and determine if they can build the reports as specified. This iterative process is intended to 'flush out' errors made during the mapping/modeling process that have left reporting areas unfulfilled. As it is found that areas of the model support the reports required those reports can be reliably built on top of the prototype model. It is noted that performance of

queries may be poor in this phase as the prototype database will not be tuned for performance. A key point we wish to make to is that the development of the Business Objects Universe and the reports required can begin PRIOR to any Informatica/DataStage mappings being written. This is a very major advantage of using SeETL as the prototype ETL tool.

7.6 Rework/Update of Models and ETL.

It is inevitable that as more reports are written there will be changes that, if implemented into the model, will make reports easier to write and faster to run. Therefore, our approach is to continue to adjust the model and the prototype ETL during the period of report construction and to delay the writing of the Informatica/DataStage ETL. Experience has shown that the sooner Informatica/DataStage ETL is written the less flexible the models become because of the high cost of rework with Informatica ETL.

7.7 Decision to 'freeze' the model.

At some point there will be a consensus decision to 'freeze the model' taken. After that point in time changes will need to be processed through a 'change control' process. Without a 'design freeze' changes tend to continue indefinitely with a diminishing return for the effort expended. Only after the model is 'frozen' should Informatica/DataStage coding begin.

7.8. Informatica/DataStage Mapping Coding.

Writing ETL Mappings in Informatica or DataStage is the most time consuming and expensive portion of all the back-end work and it is crucially important that this work is approached rigorously and methodically. SeETL can produce PDF reports for mapping specifications for Informatica/DataStage programmers. Should this be used extra information, specific to Informatica/DataStage, is added into the SeETL DesignTime workbook which is then printed on the Informatica/DataStage Specifications along with the mappings themselves. These specifications can be given to Informatica/DataStage programmers to implement. It is assumed that the Informatica/DataStage environment will be used for development even if the developers are offshore. The 'testing' of the Informatica/DataStage mappings can be 'black boxed' in that the data created by Informatica/DataStage can be tested against the data created by SeETL and if identical for all fields that should be identical the mapping can be certified as 'correct' without reviewing it in detail internally. Note that at the end of Informatica/DataStage Mapping

coding the SeETL DesignTime Workbook may or may not be used to maintain documentation of mappings for future releases of the Informatica/DataStage Mappings. It is recommended that it is used for this purpose. There is no license fee for SeETL DesignTime.

7.9 Physical Database Implementation.

Proceeding in parallel with the writing of the Informatica/DataStage Mappings and following on from the decision to freeze the data model, the DBA will begin the process of implementing the set of views that represent the model to be presented to the outside world into the best set of tables/views that can be achieved in that environment. The DBA has full authority/responsibility to implement the database in any fashion s/he chooses as long as s/he presents the same logical model as designed in the modeling process. The DBA will have been consulted during the design process for input as to 'best practices' that may be appropriate to implement.

6.10 Systems Testing.

Once the ETL has been delivered the system must be tested end-to-end to ensure that the entire system functions correctly. This will involve a number of tests including system testing, integrity testing, volumetric testing, performance testing and user acceptance testing.

8. Dimensional Design Features

8.1. Fact Tables

The first thing that we notice about star schemas is that they contain a basic central record which is stored in a 'fact table'. This input record could be a transaction record, or it could be a 'performance measures' fact record containing many performance measures for a particular period of time, usually weekly or monthly.

The BIDA fact table records are derived from the input record. The input record to be warehoused can contain any number of fields supported by your source and target database. Virtually all data types are supported. The exception is any data type that requires multiple calls to retrieve or replace the data. That is, graphics blobs etc. These things are still very rare in data warehouses. If demand requires it they will be added.

The detailed level fact table should warehouse the entire input record, including the actual fields used as real keys which are translated to integer keys, as well as the integer keys required to join the detail fact record to the dimension tables. Where no record is found in the dimension table a key of '0' is recommended on the detailed fact record.

8.2 Dimension Tables

The second thing we notice about a star schema is that there are dimension tables surrounding the central fact table. In the BIDA you can create as many dimension tables as you like.

The only real limitation around dimension tables is how many tables can be joined in a single SQL statement. With most database managers this is 16. You can have many more than 16 dimension tables joined to a fact table as part of the data model; you just cannot actually join them all to the fact table in one query.

8.3 Summary Tables

Most people building a data warehouse are familiar with the concept of summarisation and understand that summaries are necessary to reduce the cost of running the data warehouse where large fact tables are being analysed or reported against.

For example, if you are a bank with 1M transactions per day and you want 3 years of detailed transactions available you are talking about 1 billion rows in your detail level fact table. Scanning this number of rows in order to determine transaction volumes in certain branches is a complete waste of computing resources.

If you have one or more large fact tables in your data warehouse you should strive to answer the question being asked using the minimum amount of computing resources required to actually answer the question. Summaries are the #1 tool you will have available to do this.

Summary data is actually only stored in the fact tables, however, the keys which are placed on the fact table must be stored somewhere and that somewhere is recommended to be the dimension tables. To understand how summary data is stored in a dimension table consider the time dimension. It is the most common dimension and the one most easily understood.

Appendix A – Solution Scope

The BIDA solution is based on extensive practical experience and includes coverage of all real world events that occur in a telecommunications service provider organization. The BIDA solution provides the coverage needed by any fixed line, mobile or broadband operator. In addition, the solution currently has limited support for content and the data structures contained in the BIDA solution includes the following data domains.

Customer Management & Reporting	Including acquisition analysis, sales analysis, churn analysis and segmentation analysis.
Customer Integration Management	The integration and cross-referencing of customers, accounts, packages and subscriptions.
Product Management & Reporting	Including product catalog integration, product lifecycle management, product profitability analysis, and product selection and substitution analysis.
Provisioning Management & Reporting	Including service order processing analysis and aged service order reporting.
Campaign Management & Reporting	Including segment selection, response analysis, and ROI analysis for customer campaigns and communications.
Channel Management & Reporting	Including channel preferences, channel substitution and channel performance.
Supplier Management & Category Reporting (Retail)	Including supplier performance by supplier, product and organization.
Organization Performance & Reporting	Including organization performance by organization unit.
Traffic Management & Reporting (Telco)	Including un-rated, rated and re-rated CDR analysis, roaming analysis and content analysis.
Geographic Management & Reporting	Including geographic location, distribution and performance.
Switching Exchange Management & Reporting (Telco)	Including analysis of switch capability and capacity.
Transmission Network Management & Reporting (Telco)	Including analysis of network elements, capacity, performance and projections.
Service Quality Management & Reporting	Including fault analysis, complaint analysis and outage analysis.
Revenue Assurance Management & Reporting	Including all standard revenue assurance reporting templates.
Cost Management & Reporting	Including cost profiling by categories of cost, location of cost and time of cost.
Financial Management & Reporting	Including billing income analysis, profitability analysis, interconnection reporting, cash flow analysis, depreciation analysis and regulatory reporting.
Tariff Management & Reporting	Including tariff modeling and tariff re-balancing.
Credit Risk Management & Reporting	Includes customer credit risk, delinquency analysis, arrears analysis and credit scoring.
Yield Management & Reporting (Telco)	Includes post-paid and pre-paid revenue analysis.
Fraud Management & Reporting	Including propensity scoring models for fraud detection.