

DSDM and Data Warehousing

Table of Contents

1	INTRODUCTION	2
1.1	AIM	2
1.2	AUDIENCE	2
1.3	CONTRIBUTORS	2
2	WHAT IS DATA WAREHOUSING?	3
2.1	GENERIC ARCHITECTURE MODEL	3
3	WHY IS DATA WAREHOUSING DIFFERENT?	5
4	THE LIFECYCLE MODEL	6
4.1	OVERVIEW	6
4.2	PROGRAMME FEASIBILITY STUDY	6
4.3	PROGRAMME DEFINITION	6
4.3.1	<i>Business Context</i>	7
4.3.2	<i>Source Systems Inventory</i>	8
4.3.3	<i>Technical Strategy</i>	8
4.4	PROJECTS WITHIN THE PROGRAMME	9
4.5	PROJECT FEASIBILITY STUDY AND INITIATION	9
4.6	PROJECT BUSINESS STUDY	10
4.7	INCREMENT EVOLUTION	10
4.7.1	<i>Subscription Environment Evolution</i>	10
4.7.2	<i>Publication Environment Evolution</i>	11
4.8	INTEGRATION AND IMPLEMENTATION	11
4.9	DELIVERABLES	12
4.10	TOOLS AND TECHNIQUES	13
5	TEAM STRUCTURE AND ROLES	14
5.1	STRUCTURE	14
5.2	ROLES	14
6	THE DSDM PRINCIPLES	16
6.1	TESTING IS INTEGRATED THROUGHOUT THE LIFE-CYCLE	18
7	ISSUES AND RISKS	20

1 Introduction

At the time of writing, over 90% of Fortune 500 companies have data warehousing initiatives either underway or planned. Against this statistic, it is essential that there is a credible and reliable approach to data warehousing.

Experience shows that, if data warehousing is to realise the benefits which its advocates promise, then it must be approached in an incremental fashion, identifying critical business issues as the subjects for the initial projects and extending the concept outwards to other areas once the benefits have been demonstrated. Furthermore, the users of the subscription environment will typically be unclear about their specific or detailed requirements until they have had an opportunity to explore the potential of the tools and their data.

Together these factors represent a compelling case for adopting a DSDM approach to data warehousing. The findings of this White Paper underpin this viewpoint.

1.1 Aim

The objective of the White Paper is to provide advice and guidance on the effective application of Version 3 of DSDM to the development of the set of business information applications generically associated with data warehousing.

1.2 Audience

The audience for the White Paper is both business and IT project management who are considering embarking on a data warehousing project. It is assumed that readers of this White Paper will be familiar with DSDM. A brief overview of data warehousing is provided for context purposes, but for understanding the major considerations of data warehousing, the reader is encouraged to refer to other published works on the subject.

1.3 Contributors

This White Paper was put together by a Task Group reporting to the Technical Work Group. The Task Group ran several workshops and augmented the results of those workshops with written contributions from Tony Jenkins, David Lightwood and John Menzies. The Task Group would like to express its appreciation of the efforts of John Menzies in particular, who undertook the bulk of the work of producing the White Paper and preparing Task Group meeting minutes.

Stephen Godfrey	British Airways	John Menzies	Sema Group
Tony Jenkins	BCS Data Management Specialist Group	Brian Rolfe	Treasury Management Systems Ltd
Malcolm Kraushaar	Willis Corroon Group	Terri Smith	Xansa
David Lightwood	Sema Group (Task Group Chair)		

2 What is data warehousing?

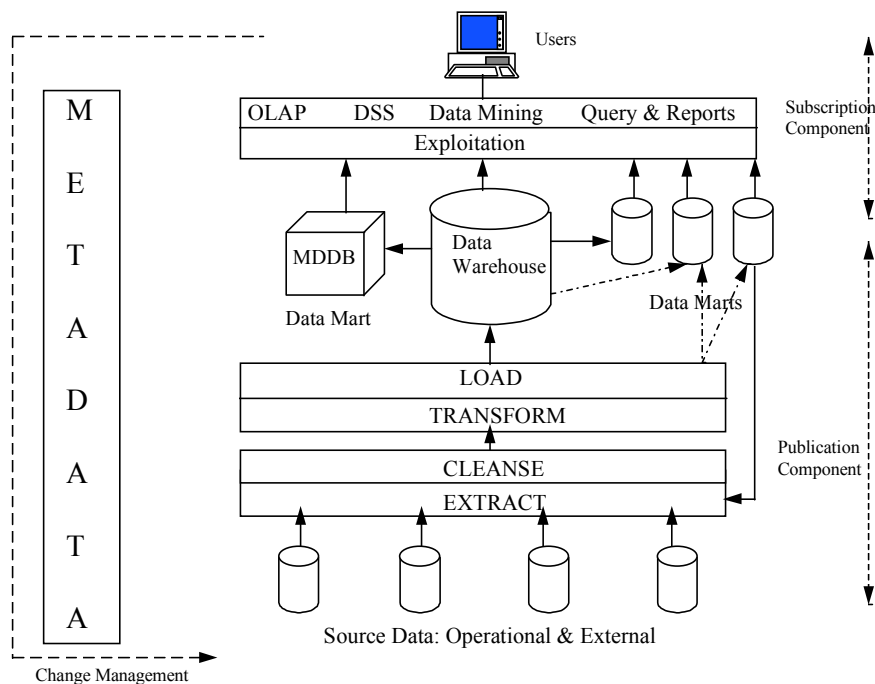
Data Warehousing is a core technology that has evolved to enable businesses to unlock the valuable information buried within their diverse operational transaction processing systems. William H Inmon, who is attributed with the original concept, provides this definition in 'What is a Data Warehouse?': "A Data Warehouse is a subject-oriented, integrated, non-volatile, time-variant collection of data organised to support management needs".

Enterprises must become more effective and efficient at what they do if they are to prosper (and even survive). Unfortunately, much of the data maintained by enterprises which would provide the insight they need in order to achieve this is locked up in databases which have evolved as discrete independent sub-systems, often designed to support a specific business process (e.g. sales order processing). These databases are unable to provide the enterprise with a consolidated view of the business or its marketplace. This is where data warehousing comes in - it provides the potential to draw together disparate data sources in such a way as to provide valuable insight into the operation of the enterprise and its interaction with its market.

In itself, the warehouse only functions as one stage of an end to end process that delivers business intelligence into the hands of the information users. Gartner Group describes business intelligence as "the process of transforming data into information, and through discovery, transforming that information into knowledge".

2.1 Generic Architecture Model

In trying to establish a common baseline definition of what a data warehouse is, the following architectural model emerged:



A significant attribute of this model (and one which is echoed repeatedly throughout this White Paper) is the concept of the split in the model between the *subscription* component and the *publication* component.

In this context the data warehouse is populated as a result of the transformation process that extracts and loads data from the operational systems. It then becomes the foundation layer of an analysis and reporting infrastructure that is capable of supporting a range of specialised information access and visualisation applications and tools. These include such familiar applications as Decision Support Systems (DSS), On-line Analytical Processing (OLAP), Executive Information Systems (EIS), Query and Reporting (Q&R) and Data Mining.

For OLAP applications, this infrastructure may include an additional multi-dimensional database containing summarised and parameterised extracts from the warehouse that enable information to be handled in terms of business related dimensions, such as time, geography, organisation, or product ranges.

Data warehouses are not things you go out and buy; they are built not bought and, if they are to succeed, must evolve in line with the changing needs of the business over time. Thus *data warehousing* is the activity associated with building, growing, sustaining and evolving data warehouses over time.

3 Why is data warehousing different?

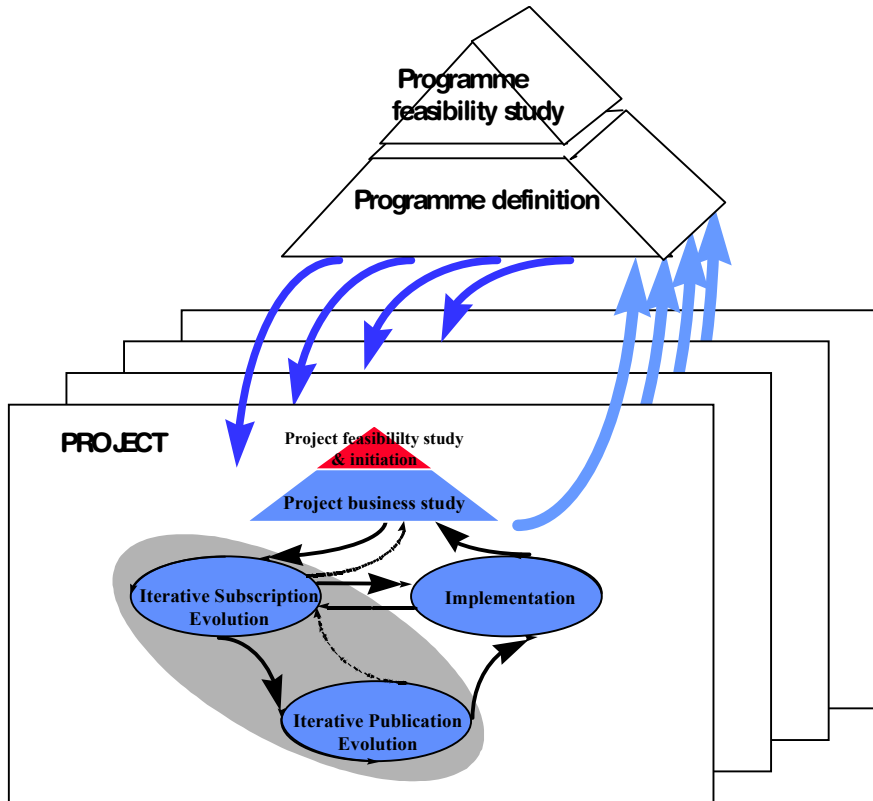
The following highlights some of the key differentiators for data warehousing:

- 1) Data warehousing has an essentially programmatic nature. A successful project will spawn further projects (each focused on a business subject) and project increments within the overall data warehouse programme.
- 2) The parallel nature of any project/increment derives from the need to deliver both the user front-end (subscription) application and the data pipeline/data store (publication). It is interesting to note the balance which must be maintained between these two environments over time, whilst individual increments may consist almost exclusively of one only. They could be solely infrastructure or exploitation increments.
- 3) Related to the preceding point is the iceberg paradigm. This refers to the anticipated balance of effort over the lifetime of the programme between the front-end (subscription) application and data pipeline/data store (publication).
- 4) Any benefit accruing from a data warehouse is entirely dependent upon the data published by other (operational) application systems. This means that data warehouse projects are essentially two-sided, with data *publishers* extracting data from operational systems and making it available to the *subscribers* who analyse the data to provide the information and insight which they seek.
- 5) This dependency upon published data also impacts data take-on in that rather than undertaking an initial data take-on exercise followed by generation of its own data, a data warehouse is in a perpetual state of data take-on, albeit of a reduced volume.
- 6) Arising directly from the dependency on published data is the need to temper high expectations with the reality of data availability.
- 7) The user roles will be twofold comprising user representatives who are responsible for data publishing (the 'operational data champions' or 'information providers') and who have intimate knowledge of the data available in the organisation, and those who represent the interests of the data subscribers who are looking to extract value and insight from the information presented (the more traditional DSDM user role).
- 8) Data warehousing is characterised by being highly data-centric. Other than the processes associated with data extract, cleansing and transformation, there is little or no process dimension to data warehousing systems. This has significant implications for the skills of the development team which will be dominated by the roles of data analyst, database designer and database administrator.

4 The lifecycle model

4.1 Overview

A macro process model emerged from the discussions. It depicts a programme control 'prism' presiding over a series of projects, one or more of which will be underway at any one time. This model is depicted in the illustration below:



The following subsections provide an outline of the aims, outputs and methods of each of the stages identified in the model. The aim is to identify and concentrate on the differences in style and focus between these stages and those of the DSDM model.

4.2 Programme Feasibility Study

This is a standard DSDM feasibility study with especial reference to the data sources, in terms of their accessibility, platforms, quality and volume.

4.3 Programme Definition

The aim of this stage is to investigate all aspects of the project to allow a full and clear definition of the scope of the proposed system. In defining the business context, performing a source systems inventory and formulating a technical strategy, the stage

will deliver the foundation for a data warehousing system that will meet the immediate requirements as well being capable of growing to meet the longer term business objectives.

The Programme Definition stage will deliver the following benefits

- Focus on the business needs ensures that the system will be designed to provide real business benefit.
- Understanding the current and future technical infrastructure enables the technical solution to be designed to integrate fully with this.
- Establishing a clear programme scope at an early stage enables controlled planning and management.
- The users' (both publishers and subscribers) involvement in the scoping and later development increases their commitment to the programme and leads to a data warehousing system that best meets their requirements.
- The Visionary's role ensures that the direction which is set is right for the organisation and is consistent with the intended objective for data warehousing within the organisation.

The key to this work is alignment, consequently the outputs from the programme definition stage are:

- 1) an analysis and statement of the *business context* and *business data context* within which the data warehousing process will operate
- 2) a *source systems inventory* to establish the potential data publishers for the data warehousing system
- 3) a *technical strategy*, aligned with the organisation's existing IT strategies and policies.

These outputs are produced in parallel and will evolve over the life of the programme. They will be revisited and probably revised over time with feedback from each separate project being used to fine tune the overall prioritised function/data list which drives the schedule of projects.

4.3.1 Business Context

This will concentrate on understanding of the current business processes in order to identify the objectives for future activity. This will include consideration of the organisation's business direction, any existing business constraints and the objectives of the data warehouse programme. There are two parallel aspects to this study:

- *Business Context* The study of which will ensure the alignment of the programme with the aims, objectives and vision of the business. As the very purpose of a data warehousing system is to inform the business and its approach, potentially causing the organisation to change the direction it takes, this is essential. Key skills include organisational and business analysis.

- *Business Data Context* It is essential to establish firmly the position and scope of the data warehousing system within the information or data architecture of the business. The warehouse must be scoped in data terms and a high level audit of corporate data quality should be used to identify any gaps and to propose how these may be plugged. Prioritisation is of the essence in this exercise, as it is data and the information which can be derived from that data, which determine the scope and scale of the data warehouse, rather than functionality per se. Key skills include data and risk analysis.

The programme definition must use the output from this work to produce the first version of the *Benefits Monitoring Strategy*. A data warehouse which fails to deliver measurable results will fail. Informed from the business and business data context this report will identify the high level goals of the programme and the critical factors which will be used to measure its achievement against those goals.

4.3.2 Source Systems Inventory

In order to establish the business data context and to identify potential sources, the work must include an audit of existing systems and their data content. This should be at a level sufficient to enable the correct prioritisation of projects/increments, rather than a complete entity level analysis of all systems. Analysis will be carried out to look at the source systems and the nature of the data. For all source systems the following will be considered and assessed:

- granularity of time
- data volumes and volatility
- data quality
- data usage
- data history
- data format (i.e. ASCII, EBCDIC, DBS/IDMSX/ISAM files, etc.)
- constraints/availability
- platforms and hardware infrastructure/connectivity.

4.3.3 Technical Strategy

This information will establish the requirements and constraints which will be used to propose potential technical solutions to meet the defined business objective. It is important that any technical solution is fully able to integrate with the current and proposed technical architecture. Current architectures will be investigated and the technical strategy will be reviewed to establish future directions.

The Technical Strategy will address:

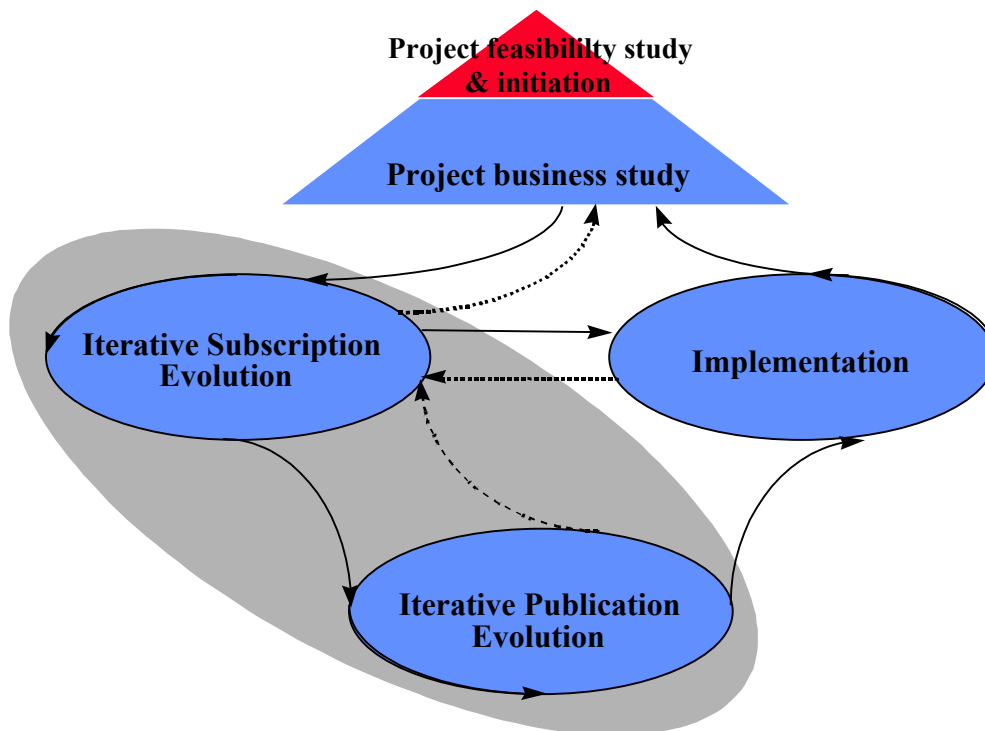
- technical constraints
- system architectures

- standards
- strategic software selection.

The scope and scale of the data warehousing programme may inform the organisation's technical strategy.

4.4 Projects within the Programme

The following figure illustrates the structure and general relationships between activities of the individual projects.



4.5 Project Feasibility Study and Initiation

This will be carried out for each project of the programme and will follow the DSDM process. Each project will have one or more subscription and publication elements. The inter-relationship between these elements needs to be considered.

The stage is essentially as in DSDM, although the work will involve a more detailed audit of the data contents of the source systems which have been identified as the data publishers for this subject area. The inventory is revisited and any issues are analysed, risks assessed and the scope focused on delivery from the outputs of this analysis.

4.6 Project Business Study

This stage will follow the DSDM process with the additional inputs for consideration of the context of the programme and any other concurrent projects. Principal deliverables will be:

- a high level data model which identifies the principal facts/business measures, dimensions and their slowly changing nature
- the augmented source systems inventory (including clarification of actual data availability)
- revisited system architecture definition (identifying new tools, new mart/fact tables and the impact on existing dimensions)
- revisited prioritised list (applying source system constraints)
- composite project plan covering both publication and subscription activities
- key non-functional requirements.

4.7 Increment Evolution

This phase often uses facilitated workshops and iterative prototyping to evolve a single data warehouse system increment. It covers the development of the data warehouse components identified in the generic data warehouse model: the subscription and publication environments, using a modified DSDM approach. An important feature of this work (and one which has a major impact on how the project is organised) is the highly parallel nature of subscription and publication environment evolution.

4.7.1 Subscription Environment Evolution

The purpose of this phase is to:

- deliver the tested subscription environment ready for integration with the data pipeline and data storage environment
- achieve functional acceptance of the subscription environment
- prepare for the roll-out of the data warehouse increment under development to the user community. The focus is on designing and developing training and user guides.

At this stage, the data storage may not have been designed or built. Consequently the basis for the prototypes will often be a temporary working database. This database may not form part of the final warehouse solution. It is there simply to facilitate development of the subscription environment which will then be integrated with the production data warehouse/data marts during the *Integration and Implementation* phase.

The principal deliverable from this phase is a fully tested and functionally accepted **subscription environment** (manifest as a set of reports/analyses and the metadata to support them) which will be integrated with the publication environment during the *Integration and Implementation* phase.

4.7.2 Publication Environment Evolution

The purpose of this phase is to establish precisely how the information requirements established in subscription environment evolution are to be met in terms of:

- the data which will deliver the required information
- the source of this data, either internal or external to the organisation and, where an item of data can be sourced from more than place, identifying which is the right source
- the transformations which must be performed on the data in order to deliver it in a form which will support the information requirements (taking full account of actual data quality as provisionally assessed during the Project Business Study)
- the structure of the data storage. Issues which will be addressed include:
 - *data marts, operational data stores* Can the requirements be met effectively from a data warehouse or are data marts or operational data stores required?
 - *scalability* The storage design must be able to accommodate the significant growth which can be expected in most data warehouse applications, both from the expansion to include further subject areas and from the 'natural' increase in data volumes which typically occurs over time.

This will be used as the basis for developing the data feed processes and for the design of the data warehouse and data marts. Also addressed during this phase is preparing operations staff to support the ongoing operation of the data publication environment and management of the warehouse.

Not every increment will include a data publication evolution. A scenario might be to deliver additional facilities within the subscription environment which are based on the data publication environment delivered in support of an earlier increment.

The principal deliverable from this phase is a fully tested, integrated and tuned **data publication environment** ready for integration with the subscription environment during the *Integration and Implementation* phase.

4.8 Integration and Implementation

There is little divergence from standard DSDM inasmuch as the data warehouse application is put into live operational use. There will be additional complexities resulting from any initial data take-on and the possible need to catch-up following the initial load (which may take several days and therefore require that those several days'/weeks' individual daily loads are sequentially loaded before the warehouse is actually 'current'). There is potentially a disproportionately large amount of effort associated with this stage associated with this data take-on activity. It is typically difficult to timebox, unless the 'fitness for business purpose' principle can be appropriately applied in the context of the age of the historical data which is taken on. In other words, it may be possible to load historical data in reverse chronological order and stop when the timebox has expired.

There are five elements in this phase:

- integrating the subscription and data publication environments to form a complete data warehouse increment
- integrating multiple subject areas (or data marts) into a single integrated data warehouse environment: this is fundamentally a consequence of the incremental approach to data warehousing system delivery
- populating the data publication environment with data (including historical data)
- testing the end-to-end process from data extract to information delivery
- tuning the performance of the end-to-end process and subscription environment to meet the non-functional requirements which were identified during the functional model iteration of the subscription environment.

The final product from this phase is a fully tested and tuned data warehousing system increment which addresses the subject area defined for the increment and is integrated with any related subject areas implemented in previous projects.

4.9 Deliverables

A number of differences between standard DSDM deliverables and those which could be expected from a data warehouse implementation were identified as follows:

- *Logical Data Model* This will probably include star or snowflake schema or cube definitions designed using appropriate decision support modelling techniques;
- *Metadata* Within the data warehousing space the 'application' is delivered as a collection of metadata which is reused by the users in various combinations, some of which were unforeseen by the developers. Similarly, when using data extraction and transformation tools, the deliverable is a collection of metadata.
- *Benefits Monitoring Strategy* Far more so than in a standard DSDM development, a data warehousing system will set out with a high level of understanding of the actual functionality which will be delivered. It could be put that the functionality is 'a set of queries/analyses' in various areas. This is unlikely to prove viable as exposure to actual data changes priorities as the possibilities become apparent. It is key to record the areas where benefits are sought rather than functionality per se. These can then be monitored during and after the completion of the prototyping phases and used to steer later increments.
- *Test Pack (plans/results)* User acceptance testing should concentrate on the 'semantic integrity' of the information presented to the user. This goes back to the composition of the actual deliverable as a set of quality assured (correct, valid) data together with the metadata components to exploit it;
- *Reviews* All reviews of prototypes must take note of the configurations of *both* the subscription and publication environments and involve both sets of users (i.e. those representing subscribers and those representing the data publishers).

4.10 Tools and techniques

Data Modelling

Data warehouses may use special database and related technologies (e.g. multi-dimensional DBMS, parallel processing architectures) and specialised forms of data modelling (e.g. star, snowflake, fact table based, denormalised) in order to efficiently handle complex queries and very large data volumes (in the gigabyte -> terabyte range). The data warehouse may be modelled as a (small) number of core multi-keyed fact tables and a number of dimension tables relating to each fact table. These modelling approaches will require an extension to the conventional ERA models more normally found in development methods.

Architectures

Special consideration must be given to architectures in data warehousing systems as there may be several distinct tools with radically different 'architectures'. Allowance must be made in the effort allowed for completion of the System Architecture Definition as decisions on tools and architectures are likely to require significantly more effort than in more 'conventional' applications.

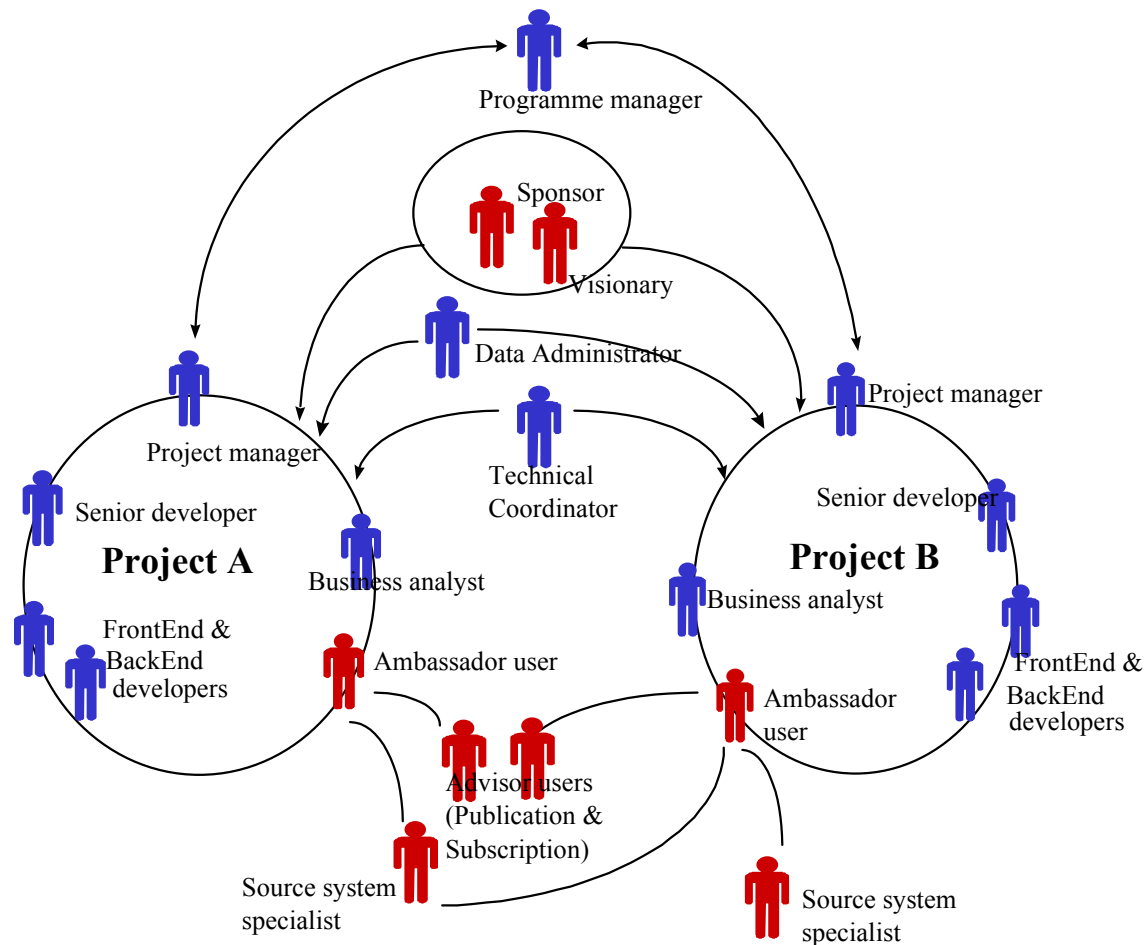
Process Modelling

Process modelling is largely redundant in the context of the front-end subscription application though it will be an important technique for the publication data pipeline application.

5 Team structure and roles

5.1 Structure

The following illustration defines the overall team structure for a data warehouse programme with concurrent project development teams.



5.2 Roles

Specific observations on various roles (relative to the standard DSDM role definition) are as follows:

- *Sponsor and Visionary* These roles may more usually relate at the level of the individual project/increment rather than the programme as a whole. This may present an issue to third parties building the warehouse for a client company.
- *Developer* In recognition of the different technologies and techniques required, different roles exist for developers of the front-end and the back-end data pipeline applications.

- *Technical Co-ordinator* The Technical Coordinator would be responsible for the overall design of the data warehouse as this will potentially publish data to a number of separate data marts and may need to 'take the strain' between competing uses (and therefore structures) for the data in the warehouse. Additionally the co-ordinator would be responsible for the technical architecture.
- *Advisor User* There will be two types of Advisor User
 - representatives of the people who will actually use the data warehousing system to glean insight into some aspect of the organisation and representatives of the data providers
 - the 'source system specialist' (this role must include representatives of those people who perform data input to the source systems to provide a view on the actual quality of the data which is collected). In live operation, this role will be responsible for the timeliness, quality and availability of the data from 'his' system.
- *Business Analyst* This role will be responsible for defining the Benefits Monitoring strategy and, once defined, for ensuring that it is realised.
- *Data Administrator* The role is responsible for the administration of the data repository - covering the data warehouse itself as well as any data marts which are generated from it.

6 The DSDM Principles

The major factor to emerge in exploring the nine principles of DSDM in relation to data warehousing is that data warehousing implementations have two distinct, but highly inter-related components: the subscription and publication 'sub-systems'. This has a persistent impact on how the DSDM principles are applied in a data warehousing context.

Active user involvement is imperative

There are two distinct groups of 'users' who must be involved in data warehouse projects:

- those representing the people who will use the data warehouse via the exploitation tools, the 'subscribers' to the data
- those who represent the 'owners' of the operational applications which are data sources to the warehouse, the data 'publishers'.

Both these groups must be represented at Ambassador User level, supported by Advisor Users as necessary.

There is perhaps less of a distinction between the Ambassador and Advisor Users. This is due to the fact that we are unlikely to be implementing a single operational business process, rather a facility to inform the decision-making process. Therefore there will be more discussion with potential for more equality of input from Advisor and Ambassador Users, although the latter will 'hold the ring' in any conflict.

There is a possible issue regarding securing the active involvement of the users who represent the source systems. The data management group or data administrator is potentially a good place to recruit from, if this exists in the organisation.

DSDM teams must be empowered to make decisions

It should be relatively easy to empower the 'subscriber' users, as they are gaining all the benefits from the warehouse. However, the 'publisher' users may need considerable authority to enact changes to the operational data capture/quality procedures to enable the warehouse to deliver higher value.

As the publishers may gain little if no direct benefit from the warehouse, a deal of tact and diplomacy will be required. This highlights the need for the Executive Sponsor role since escalation may be required if service level agreements (SLAs) are affected. In this context, SLAs relate to the operational service which is delivered by the systems which are the sources for the published data. Where there is conflict between the need to maintain an operational service according to an SLA and the need to perform data extracts to support data publishing, there may be a need for SLAs to reflect the data publishing service level required to support the data warehousing system.

The focus is on frequent delivery of products

Publication may be constrained by the development tools and strategy adopted, and by the ownership of the source data (there will be different degrees of freedom/flexibility implicit in the different tools employed). There are three scenarios:

- the model-driven approach with direct access into the source database (enabling an iterative model to be adopted)
- a specified flat file interface
- a specified flat file interface where the source system is 'owned' or run by a third party (as would be the case where the organisation had outsourced its IT systems).

There is significant scope for very rapid iteration/incremental delivery of subscription services against a warehouse whose publication services were delivered in a previous increment.

Fitness for business purpose is the essential criterion for acceptance of deliverables

The 'publication' systems may act as a constraint on the potential utility of the data warehousing system. Issues with source data must be identified as early as possible, although this is likely to be a process of discovery over time. This may represent a change lifecycle outside the increment, above the level of iterative development within the increment.

It will typically be difficult to assess 'fitness for business purpose' on day one. Confidence in 'fitness for business purpose' will increase over time as the value and quality of the information delivered by the data warehousing system is recognised. This may, to some extent, be dependent upon the progressive build-up of data over time. The Benefits Monitoring Strategy will provide a focus for monitoring fitness for business purpose.

Iterative and incremental development is necessary to converge on an accurate business solution

The 'subscriber' users are often unclear about their requirements and about the potential both of a data warehouse per se, and of the data lying 'dormant' within their operational systems. Therefore iterative and incremental development is the ideal way to implement a data warehousing solution.

A 'big bang' approach may well involve high risk and will certainly involve high costs, as well as being inimical to the early delivery/realisation of 'quick win' benefits. For this reason it is essential that the programme business study is kept at a high level and avoids architectural paralysis (whilst thinking strategically).

The relative timescales of the two development areas (publication - slower; subscription - faster) may give rise to a model in which publication and subscription developments are deliberately 'out of synch' with one preceding the other in an informed and controlled manner.

All changes during development are reversible

Publication may be less flexible than subscription, although the application of configuration management and version control regimes will mitigate the impact of any issues.

The very flexibility of the subscription environment, which may be implemented using a metadata-driven end-user toolset should be applied through an agreed release mechanism. Once the data is in the warehouse there may be a temptation to introduce new subscription functionality into the live environment 'immediately'.

Requirements are baselined at a high level

Detailed requirements for data warehousing applications typically cannot be defined up front. The subscribers often need to see the data warehousing system before their thoughts on what they need start to clarify. Therefore this principle represents the only sensible/feasible way to define requirements.

Care must be taken when scoping both the programme and the individual projects to allow a certain degree of latitude whilst working within a strategy. This is more easy to do with respect to the scoping of the subscriber functionality than that of the publication.

6.1 *Testing is integrated throughout the life-cycle*

Key points:

This is really the only way to test a data warehousing system. If all testing is left until the end of development, it is likely to be extremely difficult.

There is a significant integration activity as the live data feeds from the source systems are brought into operational use.

There may be an initial data take-on activity which may be different from the operational data feeds. Both must be tested and will potentially place different requirements on testing.

If a 'waterfall' method is used to develop the publication functionality, this will need to be addressed with an appropriate testing regime.

Testing of metadata must cover the dual role of metadata: as specification of process and as end-user guide to the contents and use of the warehouse.

Testing data warehousing systems is all about building confidence in the data. Most of the output is derived data. This makes verification more difficult as there will often be a convoluted path between the source data and the output. In this context the 'fitness for business purpose' (or good enough) is paramount if testing is not to become open-ended. Consider including an additional timebox with the specific objective of building subscriber confidence in the data as a means of accelerating the process of building confidence in the system.

A collaborative and co-operative approach between all stakeholders is essential

There are two - potentially exclusive - stakeholder groups: the publishers and the subscribers. In practice, it is likely that the publishers will put most into the project (and take significant pain in the course of doing so) with little to gain for themselves. Considerable tact and diplomacy will be required to ensure that the publishers remain on-side throughout the project. The ideal (though difficult to achieve) position is that the owners of the source systems are beneficiaries of the subscription services.

There are cultural issues around the introduction of new responsibilities for the publishers of data. The quality of data in the operational systems which feed the data warehousing system will gravitate to the level which is necessary to serve the needs of the user of those systems. This may not be good enough to meet the needs of the data warehousing system and steps will have to be taken to encourage the publishers to deliver data of the necessary quality.

7 Issues and risks

In addition to the issues highlighted in relation to each of the DSDM principles, a number of other issues or risks have been identified. These are as follows.

- The dislocation between the publication and subscription developments must be managed. It was felt that an approach of decoupling would be most appropriate, using one of a number of techniques (mock-up data, operational data store and independent, out of synch increments). Depending upon the approach adopted for the publication environment, it may be useful to refer to the DSDM Consortium White Paper on 'Managing Mixed DSDM and Waterfall Projects'.
- Time- or effort-boxing is easier to manage (and to scope) in the subscription, than the publication environment.
- The impact on prototyping with a data-centric application when data is not available at the outset. The means exist to manage this, but need recognition in developing the prototyping approach and plan.
- Use real data as soon as possible to gain confidence in the eventual solution - even if it is not cleansed. If using dummy data or sample data then there is a risk that this will be unrepresentative of the live data. It is essential to get early awareness of any existing or potential data quality or data availability issues.
- Recognise the importance of a defined release strategy - especially in the context of 'subscription increments'. Avoid uncontrolled release of untested functionality (however testing is defined). The only exception to this is in subscription-only projects where there are no changes to the publication environment (i.e. no changes to data feeds, no new data sources and no changes to the structure or content of the data warehouse).
- A data warehouse implementation does not create its own data - its success is ultimately entirely dependent upon the quality of its source data and the manner in which that is transformed for analytical use.
- The source systems may have SLAs associated with them. It is important to be aware of these and to recognise the constraints this may impose on getting access to data from these systems.
- The process of building confidence in the data warehousing system carries the risk that, as attention is focused on the data, confidence in the operational source system's data may be undermined.

- Don't under-estimate the scale of data take-on. It is often disproportionately large in relation to the project and is difficult to timebox. One possible solution is to load historical data in reverse chronological order within a timebox which is sized using MoSCoW prioritisation of historical data. There will be an absolute minimum of historical data which is needed for the data warehousing system to deliver information which is of any use at all. There will then be older historical data which should be made available if time permits and even older historical data which could be useful. There may also be historical data which is not needed to support the current project, but will be needed for subsequent projects.