Data mining, Data Warehousing and Business Intelligence Comparison among Different Development Approach

M.B. Bramarambika

Research Scholar, CMJ University, Shillong, Meghalaya

1. INTRODUCTION

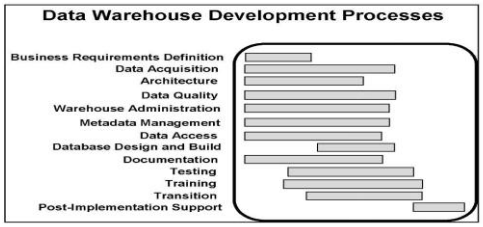
The GRT DW/BI Development Approach is a comprehensive approach to the design, development, and implementation of data warehouse solutions. It is based on input from several sources and from the direct experience of the GRT Consulting.

The GRT DW/BI Development Approach has evolved to be responsive to the dynamic nature of this business area. It is the synthesis of a detailed methodology for developing data warehouses and a methodology for doing so on an incremental basis. This generates early value for the business during the implementation process while ensuring the quality of the overall implementation effort.

The foundation is the set of Data Warehouse Development Processes, which are summarized immediately below. An overview of The Incremental Approach and then of the resulting GRT DW/BI Development Approach itself follow that discussion.

1. DATA WAREHOUSE DEVELOPMENT PROCESSES

The sections below provide perspective on each of the critical processes in a data warehouse development effort. In addition to the details discussed in each section, numerous control mechanisms and management techniques to facilitate the success of the overall project support each process.



2.1 BUSINESS REQUIREMENTS DEFINITION

The Business Requirements Definition process defines the requirements, clarifies the scope, and establishes the implementation roadmap for the data warehouse. With the direction of the client organization, strategic business goals and initiatives are established and used to direct the strategies, purpose and goals for each phase of the data warehouse solution.

Early in the process, the focus is on the enterprise aspects of the data warehouse such as enterprise information requirements, subject areas, an implementation roadmap, and a business case for the data warehouse. As the process continues, Business Requirements. Definition focuses on determining the specifics of the solution to be developed and delivered, identifying the client's information needs, and modeling the requirements.

2.2 DATA ACQUISITION

The objective of the Data Acquisition process is to identify, extract, transform, and transport the various source data necessary for the operation of the data warehouse. Data Acquisition is performed between several components of the warehouse, including operational and external data sources to data warehouse, data warehouse to data mart, and data mart to individual marts.

Early in the Data Acquisition process, data sources are identified and evaluated against the subject areas, and a gap analysis is conducted to verify that data is available to support the information requirements.

The Data Acquisition Strategy is also developed to outline the approach for extraction, transformation, and transportation of the source data to the data warehouse. The strategy includes selecting a tool or set of tools as the data pump or defining the specifications of one that must be built. If tools are to be utilized, high-level tool requirements, tool evaluations, and tool recommendations are also addressed.

As the Data Acquisition process progresses, focus shifts to the source data required to support the scope outlined in the Business Requirements Definition process. This includes the development plans for the first-time load, subsequent refresh, and sequencing of the data acquisition modules. Detailed analysis is performed on the data sources and a mapping is created between the current state of the source data and the new set of objects that define the data warehouse. With the mapping, a gap analysis is produced to validate that the information requirements can be met with the available data.

With the detailed analysis output, modules are designed and built to extract, transform, transport, and load the source data into the warehouse. Once built, the modules are tested and executed and the production database objects are populated.

2.3 ARCHITECTURE

The Architecture process specifies elements of the technical foundation and architectural design of the data warehouse. Throughout the process, the focus is on the integration of many different products and various data warehouse components to provide an extendible and scaleable architecture. The Technical Architecture, Data Warehouse Architecture and Infrastructure Roadmap are defined to outline the design and implementation of the architecture.

For the Technical Architecture, an evaluation is performed to determine whether the database environment should be distributed or centralized. Network, hardware, and software requirements are also defined and implemented, with focus on areas including acquisition, infrastructure changes, and platform configuration.

The data acquisition environment, server architecture, middleware, database sizing, and disk striping are some of the areas covered in the platform configuration. This process also establishes strategies and plans for security and control, backup and recovery, disaster recovery, and archive and restoration.

The Data Warehouse Architecture provides an integrated data warehouse environment while delivering incremental solutions. The architectural design focuses on the application of a centralized data warehouse, data marts, individual marts, metadata repositories, and incremental solution architectures. As the process continues, the development and execution of the integration plans are completed and the compliance of incremental solutions with the strategic architecture is validated.

2.4 DATA QUALITY

The objective of the Data Quality process is to provide consistent, reliable, and accurate data in the warehouse. The Data Quality Strategy is developed based upon a clear understanding of which data cleansing and integrity functions meet the needs of the customer. To facilitate timely and reliable resolution of data issues, Data Management Procedures are also defined. Early in the process, data quality tools are also evaluated and recommended.

The Data Quality process identifies the business rules for error exception handling, data cleansing, and audit and control. In addition, any variations in business rules for error handling between initial loads and subsequent updates to the data warehouse are defined. Utilizing the strategy, procedures, and tools, modules are developed to support the requirements for data quality. In order to populate the data warehouse with reliable data, the

Data Quality modules are integrated with the Data Acquisition modules to check that the quality functions are properly sequenced within the overall transformation of source data to the target environment.

2.5 WAREHOUSE ADMINISTRATION

The Warehouse Administration process specifies the strategy and requirements for the maintenance, use and ongoing updates to the data warehouse. Early in the process, the Warehouse Administration Strategy is established specifically outlining areas including version control, scheduling, data warehouse usage, security, audit and data governing. The warehouse administration workflow, tool requirements, evaluation, and testing are also addressed.

As the process continues, modules are designed and built for version control, scheduling, backup and recovery, archiving, security, audit, and data governing. In addition, several administration and monitoring tasks are addressed during the process. These include authorization to access appropriate levels of data, monitoring usage, governing queries, identifying repetitive queries, calculating metrics, defining access thresholds, adding or removing users, and updating access authority. To provide successful ongoing support and maintenance of the warehouse, this process focuses on the automation of the administration tasks, wherever possible.

2.6 METADATA MANAGEMENT

The Metadata Management process determines the Metadata Strategy, and defines requirements for metadata types, the metadata repository, metadata integration, and metadata access. The process addresses the integration of metadata for both the incremental and enterprise data warehouse solutions. A primary objective for this process is to provide technical and business views of the various aspects of warehouse metadata.

The technical view focuses on compiling metadata created during the development of the warehouse, as well as the metadata to support the management of the warehouse. The technical metadata includes:

- Data acquisition rules
- Transformation of source data to the target database
- Time and date of data
- Data authorization
- Refresh, archival, and backup schedules and results
- Data accessed, including metrics such as frequency and volume of requests For the technical staff, the access environment must support maintenance and reporting requirements in order to manage the warehouse metadata effectively.

The business view focuses on enabling end-users to understand what information is available in the warehouse and how it may be accessed. The business metadata focuses on:

- What data is in the warehouse
- How it was transformed from source to target
- The source of the data
- Information compiled while accessing the warehouse

In most cases, the selected data access tools support metadata access for end-users. Depending on the functionality of the tools, users may browse, create queries or reports, and conduct drill-down analysis on the metadata.

During this process, modules are developed for capturing, bridging, and accessing the metadata. Metadata is created by several data warehouse components, such as data acquisition, database design, and data access. Each component, especially if supported by a tool, has its own metadata storage facility and access capabilities. Therefore, the disparate metadata must be linked in this process using bridging capabilities to check consistency and to facilitate access by the appropriate personnel.

2.7 DATA ACCESS

The Data Access process focuses on the identification, selection, and design of tools that support end-user access to the warehouse data. Early in the process, a strategy is established and requirements are defined as a framework for the data access environment. Tools are evaluated, tested and recommended.

As the process continues, the user profiles are defined based on the level of data required to support their analysis, decision-making requirements, and skill level. Detailed requirements are also collected for the user interface style, queries and reports. With the user profiles in place, functional requirements, the levels of data to be accessed, and tool criteria are established for each data access component. In most cases, data access is supported by a variety of tools, rather than one tool to support every type of user.

Once tools are selected and installed, the data access objects are designed and developed including canned queries and reports, catalogs, metadata retrieval, hierarchies, and dimensions, end-user layer schemas, and user interfaces.

2.8 DATABASE DESIGN AND BUILD

The goal of the Database Design and Build process is to determine how the database objects are designed to support the data requirements and provide efficient access to the data. In addition, the logical and physical database designs are created and validated. Database designs are created for the relational and multidimensional database objects.

During this process, physical data partitioning, segmentation, and data placement is evaluated against business and user requirements versus operational constraints. In addition, indexes and key definitions are determined. This process also generates the database data definition language (DDL), and builds and implements the development, testing, and production data warehouse database objects.

2.9 DOCUMENTATION

The Documentation process centers on producing high quality textual deliverables. All user and technical documentation for the data warehouse is developed, including references, user and system operations guides, and online help.

To facilitate active and successful use of the warehouse, the Metadata Reference describes the contents of the data warehouse in business terms and provides a navigational roadmap to the contents of the data warehouse. In addition, the Warehouse Administration Reference outlines the workflow and the manual and automated administration procedures. The New Features Guide highlights any new enhancements to warehouse functionality that result from the implementation of the solution.

2.10 TESTING

The Testing process is an integrated approach to testing the quality of the various components of the data warehouse. Initially, the Testing Strategy is developed and approved by the client, followed by the creation of system, system integration, and module test plans, scripts, and scenarios. Each test is performed, including the volume tests on the physical designs for the database objects and regression testing of the enhanced warehouse against the current warehouse.

Data Acquisition Modules, Data Access functions, canned queries and reports also undergo thorough module and modules integration testing. The testing strategy must address the various components of the solution, including the ad hoc access processes.

Regression testing is performed, allowing changes to the data warehouse to be tested against a baseline, ensuring past functionality works when an enhancement is added. Volume testing is conducted on the production platform to check that performance meets the established objectives. Preparation of the acceptance environment and support for acceptance testing is also performed during the Testing process.

2.11 TRAINING

The Training process defines the development and end-user training requirements, identifies the technical and business personnel requiring training, and establishes timeframes for executing the training plans. During this process, the training plan and training materials are designed and developed, and the user and technical training is conducted.

The objective is to provide both users and administrators with adequate training to take on the tasks of operating, maintaining and using the data warehouse. Training focuses on tool training as well as the ways by which business value is generated from the information in the data warehouse. The team also trains the client's maintenance personnel and the acceptance test team.

2.12 TRANSITION

The Transition process focuses on the tasks needed to perform the cutover to the production data warehouse. It includes tasks to create the installation plan and prepare the client maintenance and production environments. During this process, the warehouse administration workflow is implemented, and the production data warehouse is available.

2.13 POST-IMPLEMENTATION SUPPORT

The Post-Implementation Support process addresses the ongoing administration of the warehouse and provides an opportunity to evaluate and review the implemented solution.

During this process, use of the data warehouse is evaluated by accessing metadata and evaluating queries and reports run against the warehouse. This information assists with the management of standard queries or reports, the end-user layer, and the identification of potential indexes.

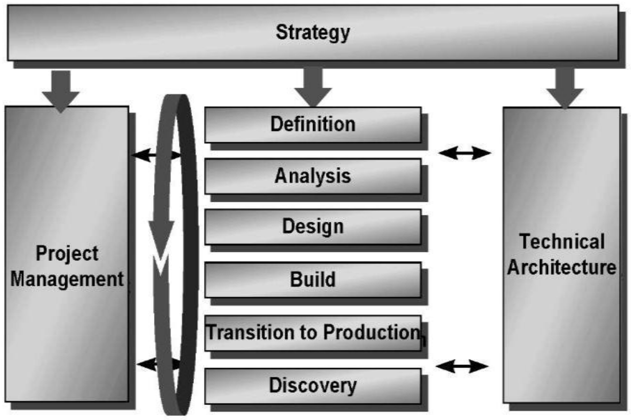
The process also focuses on several key warehouse administration functions including refreshing the warehouse, monitoring and responding to system problems, correcting errors, and conducting performance and tuning activities for the various components of the data warehouse. This includes change control for information requirements, roll-out of metadata, queries, reports, filters, and conditions, the library of shared objects, security, incorporation of new users, and the distribution of data marts and catalogs. During this process, responsibility for the data warehouse may be transferred to the client organization.

3 THE INCREMENTAL APPROACH

The Incremental Approach is an effective, proven, and preferred approach for building a data warehouse solution. In the diagram below, observe that initially an

Enterprise Strategy development effort is conducted which defines a long-term data warehouse vision from the business functionality and a technical architecture viewpoint. Following that, the incremental approach cycles through the lifecycle phases (Definition, Analysis, Design, etc.) for each increment (an increment could be a data mart or a subject area) and is supported by appropriate Project Management Methods and Enterprise Architecture Methods.

INCREMENTAL DEVELOPMENT METHODOLOGY



The Incremental Approach addresses the custom development of a data warehouse solution in a manner that generates ongoing business benefits as the effort progresses. The Incremental Approach addresses managed growth of the data warehouse through the development of incremental solutions that comply with a full-scale, enterprise data warehouse architecture.

The scoped increments are delivered in relatively short timeframes while complying with the strategic data warehouse architecture. The enterprise architecture is designed to provide a solid framework within which the long-term data warehouse can evolve.

This architecture includes the development of a central data warehouse containing corporate-wide data for various functional areas, and the functionality necessary to populate, manage, and access the full-scale data warehouse. The data warehouse also controls and feeds each data mart within the architecture. By establishing this architecture, the strategic data warehouse can grow incrementally while supporting data extensibility and avoiding an un-architected group of data marts.

The Incremental Approach begins with the Strategy phase and defines the overall data warehouse solution and architecture at a high level. During this effort, the scope of the overall solution, as well as the identification and prioritization of increments, is defined. In addition, an initial technical architecture and the data warehouse architecture are developed. Through this effort, a clear vision and scope is established for the initial incremental development effort and the implementation roadmap for the strategic data warehouse solution.

Following the phasing model, the Incremental Approach advocates developing an initial increment, which addresses a focus area and quickly provides business benefit with minimal capital outlay and minimal risk. Typically, this increment is scoped for a limited set of users, representative subset of data, and a limited infrastructure. Regardless of the scope, the solution is developed to prove critical aspects and feasibility of the total data warehouse solution, demonstrate value to the business, and act as a catalyst for further warehouse development.

Once the initial increment is finalized, subsequent increments may focus on designing business functionality or providing infrastructure functionality for the data warehouse, such as increased data content, added functionality, or the automation of warehouse administration procedures. Regardless of the focus, each increment is scoped, planned, and executed in order to deliver staged business benefit.

The addition of future increments is based on the client's business and information requirements and continues until the overall data warehouse solution is in place. Even with the development of the overall solution, a data warehouse is not considered complete or finished due to its evolutionary nature, potentially unlimited growth and required administration. Over time, the architecture is enhanced to address these factors.

At the completion of each increment, an evaluation and review are conducted during the Discovery phase. Once completed, the selection and scope of the next increment is refined. Remember, the increments should have been *defined* based on the strategy study. It is important to note that each increment follows the same phase sequence, although they may be modified based on the client's distinct needs.

4 REFERENCE

Additional insight on DW/BI methodologies can be gained by contacting GRT as indicated below or by reviewing the following publications:

1. Oracle Methods^{SM} Data Warehouse Method Handbook

2. The Data Warehouse Lifecycle Toolkit, Kimball, Reeves, Ross, Thornthwaite - Wiley Computer Publishing

3. AOracle8 Data Warehousing, Dodge, Gorman - Wiley Computer Publishing

4. Building, Using, and Managing the Data Warehouse, Barquin, Edelstein editors - Prentice Hall

5. Oracle Data Warehousing, Corey, Abbey - Osborne McGraw-Hill

6. Data Warehousing: Architecture and Implementation, Mark Humphries, Michael W. Hawkins, and Michelle C. Dy - Prentice Hall