

CLOUD APPLICATION DEVELOPMENT (GROUP 1)

PHASE 2 : ASSIGNMENT NOTEBOOK SUBMISSION

**NAME : HEMANATHAN .M**

**EMAIL : Chandruvasuki1@gmail.com**

**GitHub Repository URL :** [**https://github.com/hemanathan457/Hemanathan.git**](https://github.com/hemanathan457/Hemanathan.git)

**Title of the Project :-**

**PROJECT 7 :** Data warehouse on IBM Cloud Foundary

**Topic : Innvation**

Consider incorporating features like product reviews, wishlists, and personalized recommendations to enhance user engagement and satisfaction.

**Abstract**

Data warehousing is a collection of decision support technologies, aimed at enabling the knowledge worker to make better and faster decisions. A data warehouse is a subject-oriented, integrated, time varying, non-volatile collection of data that is used primarily in organizational decision making. Data warehouse supports on-line analytical processing, the functional and performance requirements of which are quite different from those of the on-line transaction processing applications traditionall supported by the operational databases. In this paper author suggest that for a strategic solution, the hub and spoke / centralised architecture is the more likely choice.

**Implementation:-**

Once the Planning and Design stages are complete, the project to implement the current Data Warehouse iteration can proceed quickly. Necessary hardware, software and middleware components are purchased and installed, the development and test environment is established, and the configuration management processes are implemented. Programs are developed to extract, cleanse, transform and load the source data and to periodically refresh the existing data in the Warehouse, and the programs are individually unit tested against a test database with sample source data. Metrics are captured for the load process. The metadata repository is loaded with transformational and business user metadata. Canned production reports are developed and sample ad-hoc queries are run against the test database, and the validity of the output is measured. User access to the data in the Warehouse is established. Once the programs have been developed and unit tested and the components are in place, system functionality and user acceptance testing is conducted for the complete integrated Data Warehouse system. System support processes of database security, system backup and recovery, system disaster recovery, and data archiving are implemented and tested as the system is prepared for deployment. The final step is to conduct the Production Readiness Review prior to transitioning the Data Warehouse system into production. During this review, the system is evaluated for acceptance by the customer organization.

**Architecture Review and Design:-**

The Architecture is the logical and physical foundation on which the Data Warehouse will be built. The Architecture Review and Design stage, as the name implies, is both a requirements analysis and a gap analysis activity. It is important to assess what pieces of the architecture already exist in the organization (and in what form) and to assess what pieces are missing which are needed to build the complete Data Warehouse architecture.

During the Architecture Review and Design stage, the logical Data Warehouse architecture is developed. The logical architecture is a configuration map of the necessary data stores that make up the Warehouse; it includes a central Enterprise Data Store, an optional Operational Data Store, one or more (optional) individual business area Data Marts, and one or more Metadata stores. In the metadata store(s) are two different kinds of metadata that catalog reference information about the primary data.

Once the logical configuration is defined, the Data, Application, Technical and Support Architectures are designed to physically implement it. Requirements of these four architectures are carefully analyzed so that the Data Warehouse can be optimized to serve the users. Gap analysis is conducted to determine which components of each architecture already exist in the organization and can be reused, and which components must be developed (or purchased) and configured for the Data Warehouse.

The Data Architecture organizes the sources and stores of business information and defines the quality and management standards for data and metadata.

The Application Architecture is the software framework that guides the overall implementation of business functionality within the Warehouse environment; it controls the movement of data from source to user, including the functions of data extraction, data cleansing, data transformation, data loading, data refresh, and data access (reporting, querying).

The Technical Architecture provides the underlying computing infrastructure that enables the data and application architectures. It includes platform/server, network, communications and connectivity hardware/software/middleware, DBMS, client/server 2-tier vs.3-tier approach, and end-user workstation hardware/software. Technical architecture design must address the requirements of scalability, capacity and volume handling (including sizing and partitioning of tables), performance, availability, stability, chargeback, and security.

**Tool Selection:-**

The purpose of this stage is to identify the candidate tools for developing and implementing the Data Warehouse data and application architectures, and for performing technical and support architecture functions where appropriate. Select the candidate tools that best meet the business and technical requirements as defined by the Data Warehouse architecture, and recommend the selections to the customer organization. Procure the tools upon approval from the organization.

It is important to note that the process of selecting tools is often dependent on the existing technical infrastructure of the organization. Many organizations feel strongly for various reasons about using tools for the Data Warehouse applications that they already have in their "arsenal" and are reluctant to purchase new application packages. It is recommended that a thorough evaluation of existing tools and the feasibility of their reuse be done in the context of all tool evaluation activities. In some cases, existing tools can be form-fitted to the Data Warehouse; in other cases, the customer organization may need to be convinced that new tools would better serve their needs.

It may even be feasible that this series of activities is skipped altogether, if the organization is insistent that particular tools be used (no room for negotiation), or if tools have already been assessed and selected in anticipation of the Data Warehouse project.

Tools may be categorized according to the following data, technical, application, or support functions:

- Source Data Extraction and Transformation

- Data Cleansing

- Data Load

- Data Refresh

- Data Access

- Security Enforcement

- Version Control/Configuration Management

- Backup and Recovery

- Disaster Recovery

- Performance Monitoring

- Database Management

- Platform

- Data Modeling

- Metadata Management

**BENEFITS OF DATA WAREHOUSE METHODS**

With data warehousing, you can provide a common data model for different interest areas regardless of data's source. In this way, it becomes easier to report and analyze information.

Find out inconsistencies and resolved before loading of information in data warehousing, this makes the reporting and analyzing process simpler.

The best part of data warehousing is that the information is under the control of users, so that in case the system gets purged over time, information can be easily and safely stored for longer time period.

Because of being different from operational systems, a data warehouse helps in retrieving data without slowing down the operational system.

Data warehousing enhances the operational business applications values and customer relationship management systems.

Data warehousing give a way to proper functioning of support system applications like trend reports, exception reports and the actual performance analyzing reports.

Precisely, a data warehouse system proves to be helpful in providing collective information to all its users. It is mainly created to support different analysis, queries that need extensive searching on a larger scale.

Functions of Data Warehouse Tools and Utilities

The following are the functions of data warehouse tools and utilities −

Data Extraction − Involves gathering data from multiple heterogeneous sources.

Data Cleaning − Involves finding and correcting the errors in data.

Data Transformation − Involves converting the data from legacy format to warehouse format.

Data Loading − Involves sorting, summarizing, consolidating, checking integrity, and building indices and partitions.

Refreshing − Involves updating from data sources to warehouse.

Update-Driven Approach

This is an alternative to the traditional approach. Today's data warehouse systems follow update-driven approach rather than the traditional approach discussed earlier. In update-driven approach, the information from multiple heterogeneous sources are integrated in advance and are stored in a warehouse. This information is available for direct querying and analysis.

**The Data Warehouse Process :-**

The james martin + co Data Warehouse Process does not encompass the analysis and identification of organizational value streams, strategic initiatives, and related business goals, but it is a prescription for achieving such goals through a specific architecture. The Process is conducted in an iterative fashion after the initial business requirements and architectural foundations have been developed with the emphasis on populating the Data Warehouse with "chunks" of functional subject-area information each iteration. The Process guides the development team through identifying the business requirements, developing the business plan and Warehouse solution to business requirements, and implementing the configuration, technical, and application architecture for the overall Data Warehouse. It then specifies the iterative activities for the cyclical planning, design, construction, and deployment of each population project. The following is a description of each stage in the Data Warehouse Process. (Note: The Data Warehouse Process also includes conventional project management, startup, and wrap-up activities which are detailed in the Plan, Activate, Control and End stages, not described here.)

**ADVANTAGES:-**

This approach has the following advantages −

This approach provide high performance.

The data is copied, processed, integrated, annotated, summarized and restructured in semantic data store in advance.

Query processing does not require an interface to process data at local sources.

**DISADVANTAGES:-**

Query-driven approach needs complex integration and filtering processes.

This approach is very inefficient.

It is very expensive for frequent queries.

This approach is also very expensive for queries that require aggregations.

**CONCLUSION:-**

An independent data mart is more likely to be selected if resources are limited. It is observed that in the key factors influencing the choice of a bus architecture is a high need to share data or information between departments. Where as in Hub and Spoke / Centralised architecture tends to be selected where the data warehouse is considered to be an integral part of a strategic solution hence there is a high need for data to be made freely available between business units. While the need for information to be made freely available between business units is important in the choice of either of these architectures, it is the bus architecture that will tend to be chosen. Similarly, if the data warehouse is required quickly the bus architecture has proved popular. However, as a strategic solution, it is the hub and spoke / centralised architecture that is the more likely choice.

**GitHub Repository URL :** [**https://github.com/hemanathan457/Hemanathan.git**](https://github.com/hemanathan457/Hemanathan.git)