Modeling a Data Integrator in SQL Server Integration Services

Using SSIS in GARMS

[1]Parul Purohit [2]Neha Purohit
[1, 2]Acropolis Institute of Technology and research Indore, India

Abstract— *This report presents a Data Integrator model for different legacy database systems. A Data Integrator tool is a Microsoft SQL Integration Services package (also known as MS SSIS) that locates data residing inside Microsoft Excel Sheets and updates the central Application database System of GARMS (Global Assessment for Response Management).Currently there are three Microsoft Excel Sheets that contains the legacy Application data as* ***(i)GARMS_legacyProgramInfo_2.0.xls (ii)GARMS_LegacyRespondentInfo_2.0.xls (iii)GARMS_LegacyResponseInfo_2.0.xls.*** *This is often referred to an Intermittent Data Store in the context of Data Integrator tool. Different legacy applications (hereby known as source databases) containing data should be migrated to the new application (hereby known as target database) as GARMS Application. These source databases are either in IBM Lotus Notes, MS Access or legacy .Net Applications. The data from these source databases is extracted by their respective DBA and saved inside one or multiple excel sheets. For the purpose of discussion these three Microsoft Excel Sheets will be referred to as Intermittent Data Stores. There should be a proper format and structure of the Intermittent Data Store. This process is time consuming and prone to errors. The current Data Integrator tool performs data migration from these Intermittent Data Stores into Application database (GARMS). A Data Integrator Tool run as a scheduled job on Microsoft SQL Server 2005.This Data Integrator Tool reads, validates, transforms the data in sequence of Excel sheet data format by applying business Rules of Application, Database on the data and finally saves in Database.*

Here we will write a VB Script for a Data Integrator Model to define all the Business Logic for GARMS Application and describe how the system will work together. We also measure the performance of the system using various parameters such as the time taken to complete a job of migrating a data from Excel

Sheet Files to GARMS Application Database, as well as to describe different atomic model of our

Data Integrator Model. We also examine current scenarios, assumptions of our model, and discuss future enhancements.

I. INTRODUCTION

Business intelligence (**BI**) refers to computer-based techniques used in spotting, digging-out, and analyzing business data, such as sales revenue by products or departments, or by associated costs and incomes. BI technologies provide historical, current, and predictive views of business operations. Common functions of business intelligence technologies are reporting, online analytical processing, analytics, data mining, business performance management, benchmarking, text mining, and predictive analytics.BI is a generic term to describe leveraging the organization's internal and external information assets for making better business decisions.

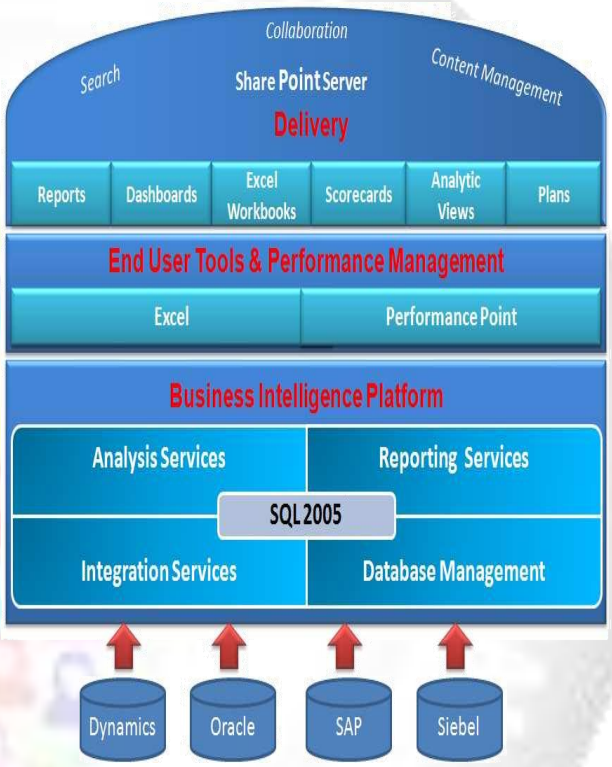


Fig.1: MicrosoftBiPlatform

II. SQLSERVER INTEGRATION SERVICES

(SSIS) is a component of the Microsoft SQL Server database software which can be used to perform a broad range of data migration tasks.

SSIS is a platform for data integration and workflow applications. It features a fast and flexible data warehousing tool used for data extraction, transformation, and loading (ETL). The tool may also be used to automate maintenance of SQL Server databases and updates to multidimensional cube data. First released with Microsoft SQL Server 2005, SSIS replaced Data Transformation Services, which had been a feature of SQL Server since Version 7.0. Unlike DTS, which was included in all versions, SSIS is only available in the "Standard" and "Enterprise" editions.

A. SSIS provides following features which are as under:-

1) Connections: -A connection includes the information necessary to connect to a particular data source. Tasks can

reference the connection by its name, allowing the details of the connection to be changed or configured at runtime.

2) Tasks: - *A*task is an atomic work unit that performs some action. There are a couple of dozen tasks that ship in the box, ranging from the file system task (which can copy or move files) to the data transformation task. The datatransformation task actually copies data; it implements the ETL features of the product.

3) Precedence constraints: -Tasks are linked by precedence constraints. The precedence constraint preceding a particular task must be met before that task executes. The runtime supports executing task in parallels if their precedence constraints so allow. Constraint may otherwise allow different paths of execution depending on the success or failure of other tasks. Together with the tasks, precedence constraints comprise the workflow of the package.

4) Event handlers: -A workflow can be designed for a number of events in the different scopes where they might occur. In this way, tasks may be executed in response to happenings within the package — such as cleaning up after errors.

5) Variables:-Tasks may reference variables to store results, make decisions, or affect their configuration. A package may be saved to a file or to a store with a hierarchical namespace within a SQL Server instance. In either case, the package content is persisted inXML. Once completed, the designer also allows the user to start the package's execution. Once started, the package may be readily debugged or monitored.

B. Features ofthe Data Flow Task in SSIS Are As Under:-

SSIS provides the following built-in transformations:

1. Conditional Split
2. Multicast
3. Union-All, Merge, and Merge Join
4. Sort
5. Fuzzy Grouping
6. Lookup and Fuzzy Lookup
7. Percentage Sampling and Row Sampling
8. Copy/Map, Data Conversion, and Derived Column
9. Aggregation
10. Data Mining Model Training, Data Mining Query, Partition Processing, and Dimension

11Pivot and

12. Slowly Changing Dimension
13. Script Component

III. SQL SERVER REPORTING SERVICES

SQL Server Reporting Services (SSRS) is a server-based report generation software system from Microsoft. It can be used to prepare and deliver a variety of interactive and printed reports. It is administered via a web interface. Reporting services features a web services interface to support the development of custom reporting applications. SSRS competes with Crystal Reports and other business intelligence tools, and is included in Developer, Standard, and Enterprise editions of Microsoft SQL Server as an install option. Reporting Services was first released in 2004 as an add-on to SQL Server 2000. The second version was released as a part of SQL Server 2005 in November 2005. The latest version was released as part of SQL Server 2008 in August 2008.

Reports are defined in Report Definition Language (RDL), an XML markup language. Reports can be designed using recent versions of Microsoft Visual Studio, with the included Business Intelligence Projects plug-in installed or with the included Report Builder, simplified tool that does not offer all the functionality of Visual Studio. Reports defined by RDL can be generated in a variety of formats including Excel, PDF, CSV, XML, TIFF (and other image formats), and HTML Web Archive. SQL Server 2008 SSRS can also prepare reports in Microsoft Word (DOC) format.

IV. SQLSERVERANALYSIS SERVICES

Microsoft released Analysis Services 2000. It was renamed from "OLAP Services" due to the inclusion of data mining services. Analysis Services 2000 was considered an evolutionary release, since it was built on the same architecture asOLAP Services and was therefore backward compatible with it. Major improvements included more flexibility in dimension design through support of parent child dimensions, changing dimensions, and virtual dimensions. Another feature was a greatly enhanced calculation engine with support for unary operators, custom rollups, and cell calculations. Other features were dimension security, distinct count, and connectivity over HTTP, session cubes, grouping levels, and many others. In 2005, Microsoft released the next generation of OLAP and data mining technology as Analysis Services 2005. It maintained backward compatibility on the API level: although applications written with OLE DB for OLAP and MDX continued to work, the architecture of the product was completely different. The major change came to the model in the form of UDM - Unified Dimensional Model. Microsoft SQL Server 2008 helps enable organizations to build comprehensive, enterprise-scale analytic solutions that deliver actionable insights through familiar tools. New Features by SQL Server Analysis Services 2008.

1. Develop solutions quickly with the new, streamlined Cube Designer.
2. Takeadvantage of enhanced Dimension and Aggregation Designers.
3. Create attribute relationships easily by using the new Attribute Relationship Designer.
4. Avoid common design problems by using best practice.
5. Optimizeperformancewith subspace computations.
6. Enablehigh-performance “what if” scenarios by using MOLAP enabled write-back.
7. Take advantage of enhanced data mining structures and improved Time Series support.
8. Monitor and optimize analytical solutions by using analysis.

V. DESCRIPTION OF THE PROJECTS:

A. GARMS: -(Global Assessment for Response Managementsystem)

The Global Assessment of Service Quality (ASQ) Team, a business unit within GAIBD, offers a unique service to Global Account Teams and individual member firms in the form of a survey management support programme.

GARMS application was designed around the core business process of ASQ team to enable global and consistent delivery of their services within and outside E&Y. For more information just go through the Scope Document of GARMS project included in a given folder.

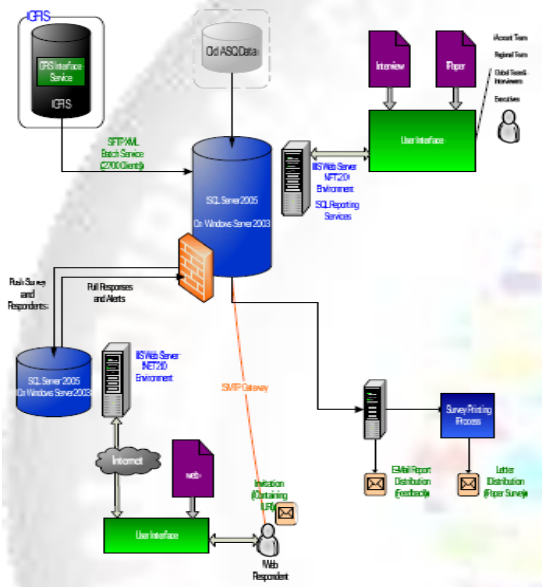


Fig4 SOLUTION CONTEXT DIAGRAM FOR GARMS PROJECT

B. GFIS:-

(Global Financial Information System)

The system is expected to interface and read data from GFIS i.e. Global Financial Information System. The following attributes are expected to Be read:

1. Global Ultimate DUNS
2. Client Name
3. Client Head Office
4. Area Code (EY)
5. Area Description (EY)
6. Channel Number (EY)
7. Industry [Sector] Description (EY)
8. Industry [Sector] Code (EY)
9. Industry Group Description (EY)
10. Industry Group Code (EY)
11. Global Market Segment (EY)
12. GCSP GPN# (EY) e.g. US00022974
13. GCSP Name (EY)
14. GCSP Business Unit (EY)
15. GCSP Office (EY)
16. Last Financial Year Revenue (EY)
17. Service Line

GFIS Feed is a set of 9 XML based files that Contain GFIS data to be merged in the GARMS Application database. These XML files are Mentioned as follows:

1. Client Information
2. Service Line
3. Market Segment
4. Country
5. Channel
6. Industry Sector
7. Area
8. Account Country Wise revenue
9. Account Service Line wise revenue

The below diagram illustrates the GFIS Feed and its components. As seen in the diagram, there are 2 distinct components as shown in following diagram:

1. Copying GFIS Feed XML Files from GFIS Server to GARMS Application / Database Server.
2. Read the XML files, Validate the XML data using pre-defined business rules and process the data in GARMS MS SQL Server 2005 Database. For more information just go through the Scope Document of GARMS project included in a given folder.

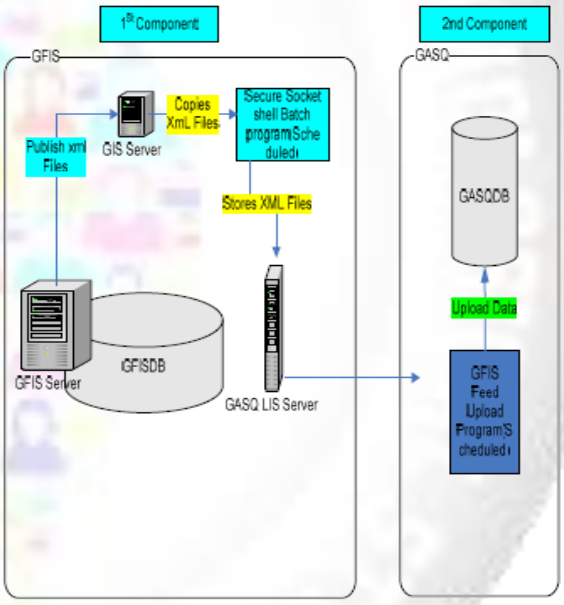


Fig5 CONTEXT DIAGRAM FOR GFIS PROJECT

VI. WORKFLOW OF THE DATA INTEGRATOR MODEL

GARMS Application is a Survey Management software application that is extensively used by ASQ Team to launch surveys with E&Y clients & monitor its responses and analyze the same with an objective of improving the relationship / engagement with its esteem clientele. GARMS Data Integrator tool is a MS SQL Integration

Services package (also known as MS SSIS) that locates data residing inside MS Excel Sheets and updates the central GARMS Application database. Currently there are three different MS Excel Sheets that contain the legacy ASQ Application data. The three different MS Excel Sheets are named as

GARMS_legacyProgramInfo_2.0.xls,

GARMS_LegacyRespondentInfo_2.0.xls and GARMS_LegacyResponseInfo_2.0.xls that also indicates the type of data stored in the MS Excel Sheets. This is often referred to an

Intermittent Data Store in the context of GARMS Data Integrator tool. E&Y has different legacy ASQ applications (hereby known as source databases) containing survey data that E&Y wants to migrated to the new GARMS application (hereby known as target database). These source databases are either in IBM Lotus Notes, MS Access or legacy .Net Applications. The data from these source databases is extracted by their respective DBA and saved inside three different MS Excel Sheets viz: GARMS_legacyProgramInfo_2.0.xls,

GARMS_LegacyRespondentInfo_2.0.xls and

GARMS_LegacyResponseInfo_2.0.xls. For the purpose of discussion these three MS Excel Sheets will be referred to as Intermittent Data Stores. The format and structure of the Intermittent Data Store is agreed between E&Y and HTL. We understand from our discussions with E&Y that this process is time consuming and prone to errors.The current GARMS Data Integrator tool performs data migration from these Intermittent Data Stores into GARMS Application database.

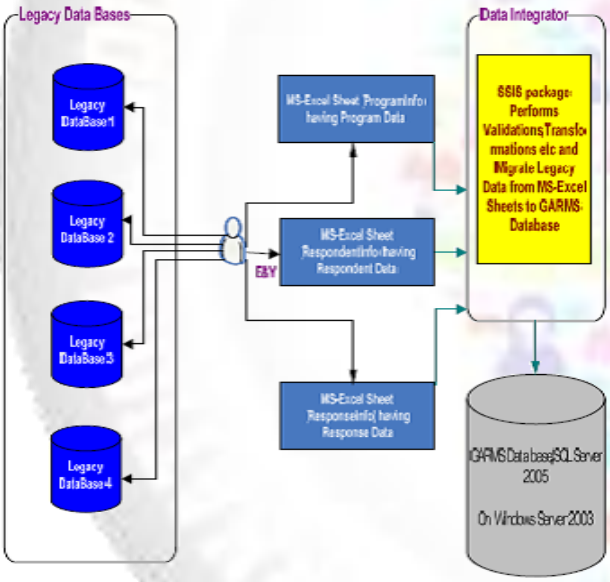


Fig6 GARMS DATA INTEGRATOR MODEL

VII. WORKING OF THE DATA INTEGRATOR TOOL

The following flow chart illustrates the schematic working of the current Data Integrator Tool

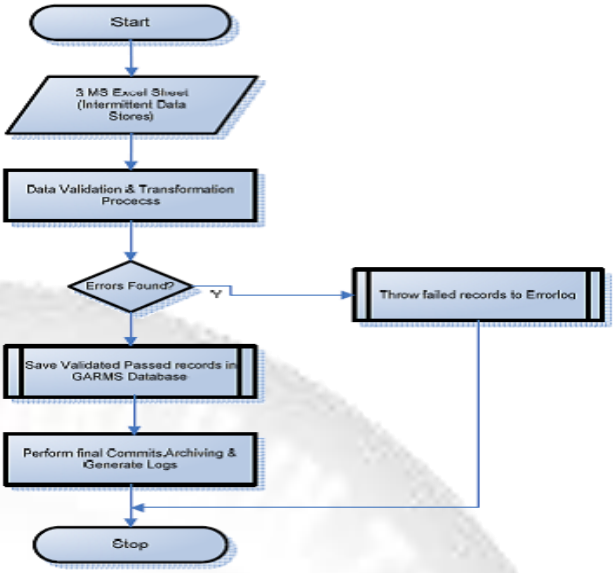
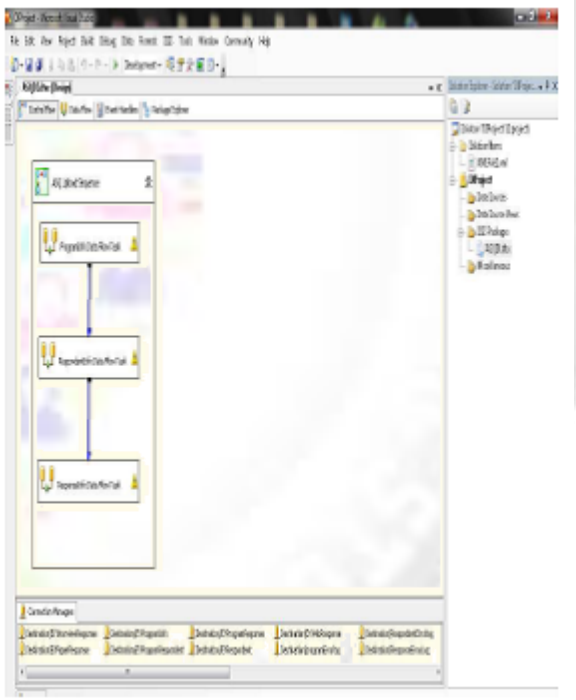


Fig7 WORKING OF A DI TOOL

VIII. SCREEN SHOTS FOR DI TOOL CONTROL FLOW DIAGRAM FOR DI TOOL



Atomic Models for Data Integrator Tool:

1. It is expected that E&Y will put the threeIntermittent Data Stores in the predefinedlocation on QA \ Production server on whichGARMS Data Integrator Tool is deployed.
2. GARMS Data Integrator Tool run as a scheduled job on MS SQL Server 2005 as perfrequency defined by E&Y. GARMS Data Integrator Tool reads, validates, transforms thedata in sequence of ProgramInfo, RespondentInfo, ResponseInfo by applying business Rules of GARMS Application / GARMS Database on the data and finally savesin GARMS Database,

3. Program Info :
   a. GARMS Data Integrator Tool verifies whether the Program data to be uploaded is already exists in the database or not, If not exits then performs the checks as per the 'Validations & Business Rules' mentioned in the rules sheet of ProgramInfo Template i.e. in the GARMS_legacyProgramInfo_2.0.xls.
   b. The valid, transformed data which confirms to GARMS application/Database business rules and Integrity will be migrated into GARMS Database.
4. RespondentInfo :
   a. GARMS Data Integrator Tool verifies whether the Respondent Data to be uploaded is already exists in the database or not, If not exits then performs the checks as per the 'Validations & Business Rules' mentioned in the rules sheet of RespondentInfo Template i.e. GARMS_LegacyRespondentInfo_2.0.xls.
   b. The valid, transformed data which confirms to GARMS application/Database business rules and Integrity will be migrated into GARMS
5. ResponseInfo :
   a. GARMS Data Integrator Tool verifies whether the Response Data to be uploaded is already exists in the Database or not, If not exits then performs the checks as per the 'Validations& Business Rules 'mentioned in the rules sheet of Response Info Template i.e. GARMS_LegacyResponseInfo_2.0.Xls.
   b. The valid, transformed data which confirms to GARMS application/Database business rules and Integrity will be migrated into GARMS Database.
6. Once the above data migration is completed,GARMS Data Integrator Tool performs FinalCommits as Consolidated insertions, updationsfor all uploaded records.
7. All the data which failed in the above process is shifted to Error Database.
8. At the end GARMS Data Integrator Tool performs archiving i.e. maintaining history of uploaded Intermittent Data Stores, and Generate logs files in predefined location on QA \ Production Server such as creating error logs from Error Database with cause of rejection, date and other details.

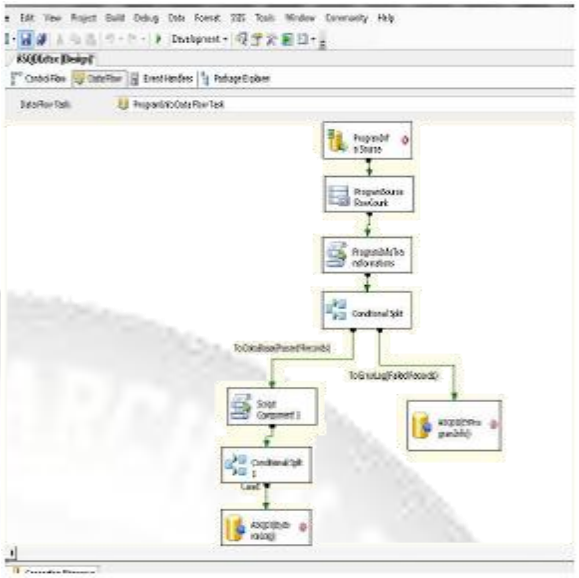


Fig.8: ProgramInfo Data Flow task in DI Tool

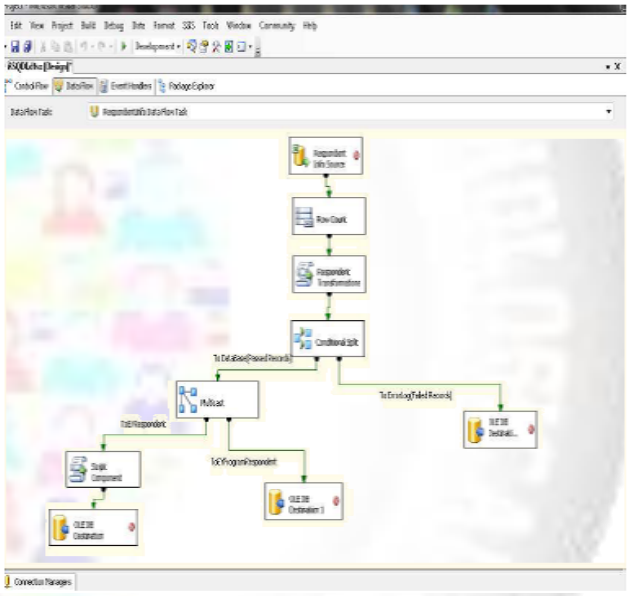


Fig.9: RespondentInfo Data Flow task in DI Tool

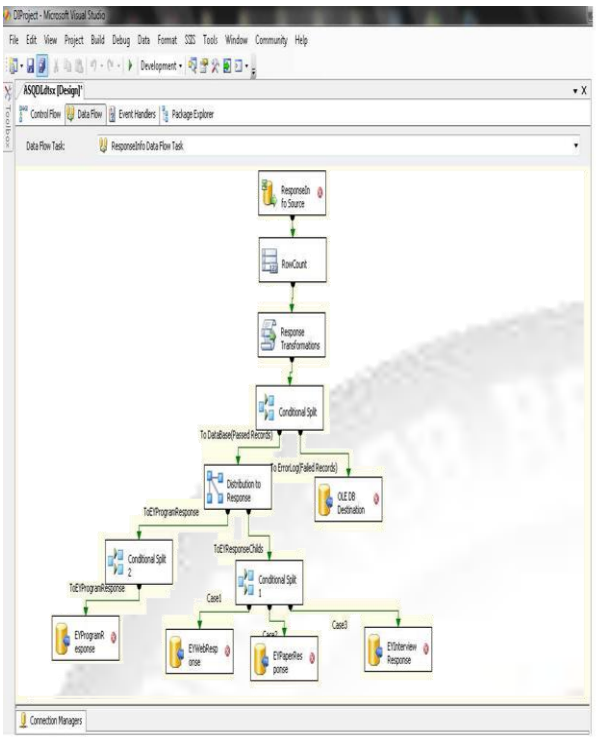
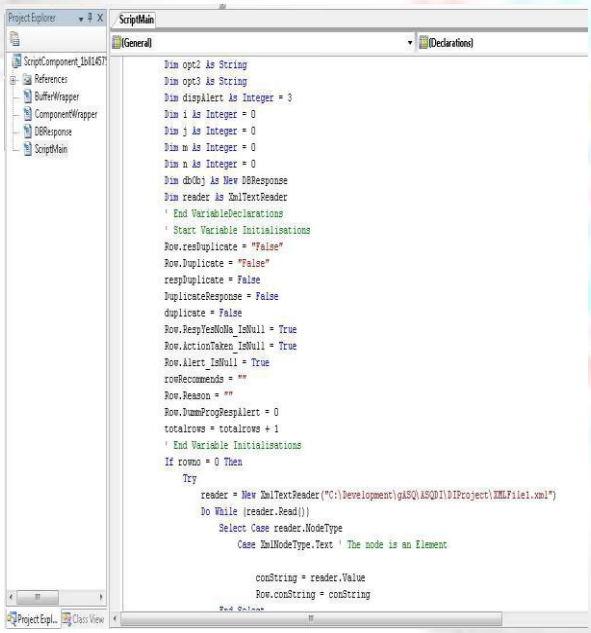


Fig.10:Response Info Data Flow task in DI Tool

IX. RESULTS AND CONCLUSION:

Screenshot of the Visual Basic Scripting in DITool in Response Info



A. With the Microsoft Excel format, specific versions of Excel have a maximum number of rows that can be displayed. Pre-version 8.0 (pre-Excel '97) exports a maximum of 16,384 records or rows in 15 minutes. Microsoft Excel 8.0 (97), Excel 2002 and Excel 2003 have a limit of 65,536 rows. Other limitations are mentioned as follows:

1) There can be maximum 256 columns and each column can contain data up to 255 characters only.
2) Length of cell contents (text): 32,767 characters. Only 1,024 display in a cell; all 32,767 display in the formula bar.

B. Enhanced GARMS Data Integrator tool will support only MS Excel 2003. The template for importing data into GARMS Application using Enhanced GARMS Data Integrator tool will be provided by HTL to E&Y for review and approval. We tested the system performance time or the processingtime which is around 15 minutes for a Data integrator tool.

The following reports are designed and developed usingMS SQL Server Reporting Services 2005 after SSIS migration of data.

1. Executive Report
2. ASQ INPUT TO GPAL
3. ASQ INPUT TO APAL
4. Area Report
5. CSP Report
6. ALERT-Area
7. ALERT-Master
8. ALERT-Sector
9. ALERT-Service Line
10. Service Line Priority Report
11. Service Line Report
12. Industry Sector Report

These reports are viewed as standalone using MS Reporting Services 2005 Interface and not integrated into the GUI of GARMS application.

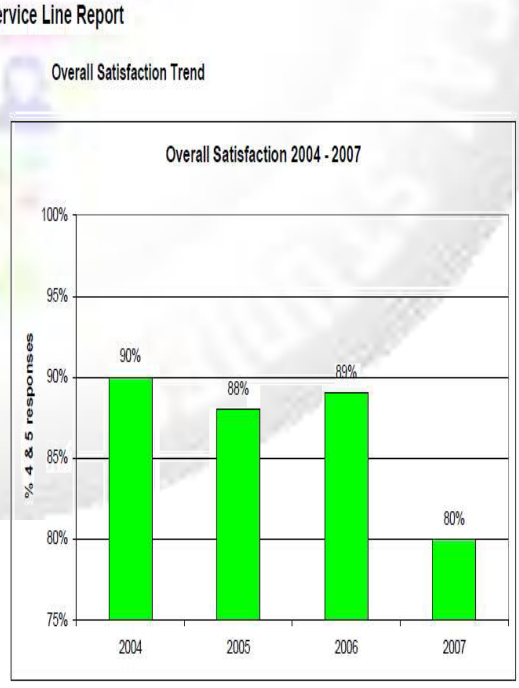


Fig.11: Service Line Report generation in GARMS Application using SQL Server Reporting Services

X. LIMITATIONS AND FUTURE WORK:

The current constrains are not in the GARMS Data Integrator tool but it is in the process that is followed by E&Y to extract the legacy data from legacy ASQ applications that are currently in different technologies to saving data in the 3 different MS Excel Sheet i.e. Intermittent Data Store is very time consuming and prone to errors. Future works on improving the system can address these limitations. The proposed enhancements to GARMS Data Integrator tool is made with an objective to simplify the process of saving data in the Intermittent Data Stores. E&Y expects the three MS Excel Sheets viz:

GARMS_legacyProgramInfo_2.0.xls,

GARMS_LegacyRespondentInfo_2.0.xlsand GARMS_LegacyResponseInfo_2.0.xls to be consolidated into a single MS Excel Sheet. All the Business Rules and Technical Parameters remain the same as per the current GARMS Data Integrator tool.

REFERENCES:

[1] The Realization of Integration for Business Intelligence Between SQL Server 2005 Reporting Services office share Point Server 2007 by Dongyun Wang 2010
[2] http://en.wikipedia.org/wiki/SQL_Server_Integration_Services
[3] http://msdn.microsoft.com/enus/library/ms141026.aspx
[4] Application of SQL Server in Data Mining Zhansheng Zhang ; Guicheng Wang ; Lei Yang 2010