Introduction

Many organizations have distributed systems where data required to make business decisions come from multiple sources. Different teams consuming this data come up with their own business logic to determine KPIs. This leads to volumes of reports which leverage fragmented data and no centralized conformation, making it very hard for users to reconcile data outside of their teams or business division and many wasted hours investigating differences in reports. Today, with petabytes of operational and customer’s data it becomes imperative to have solutions to replace the existing manual data consolidation, manual business rule application and manual data process management. There is also wide recognition that a proven strategy for consolidated reporting is to provide the ability to have a single unified data store with one version of the truth for all businesses and customers’ data needs. In today’s market of increasing competition there is push and a need to get business intelligence projects done faster, more efficiently, and cost effectively. For any large-scale Enterprise Data Warehouse or business intelligence application, there is often waste and redundancy in the Extract Transform and Load (ETL) process, making it one of the most expensive and time consuming components to build. This white paper explains how iSoftStone addresses these issues with RapidStart® Business Intelligence (BI) Solution Accelerator, which incorporates best practices of ETL development while facilitating a more efficient development process in conjunction with Microsoft SQL Server Integration Services (SSIS).

Problem Statement

Below are some of the core challenges of a typical ETL process.

- Onboarding a data source takes significant time as it requires users to reconcile foreign schemas with complex existing systems.
- Reusability of the code is difficult between data sources.
- Metadata management becomes increasingly difficult with each new dimension or source system added.
- High maintenance cost when significant changes happen in the data sources.
- Business rules are applied in multiple areas – ETL, local database, files, reports requiring more time for reconciling and validating disparate reports when discrepancies occur.
- No centralized location to track and monitor ETL jobs running on multiple servers.
- No standardization for package creation. When using SSIS anyone can create packages leading to duplicates and potentially a management nightmare.
- Managing ETL processes, dependencies and job logic is complicated and time consuming. Building a management infrastructure to support ETL processing is complex.
RapidStart BI Solution Accelerator

RapidStart BI Solution Accelerator is a high value, high impact, customizable BI data ETL framework that enables dynamic creation and execution of SSIS ETL packages across wide variety of data sources. It uses a structured approach and methodology to dynamically create and run parallel executions of SSIS ETL packages. This framework facilitates a more efficient development process in conjunction with SSIS and adds to its features, enhancing the capture and reusability of ETL metadata which is stored and maintained in SQL Server database tables.

RapidStart BI Solution Accelerator includes iSoftStone’s pre-built code that can be leveraged across a variety of application scenarios to provide a consistent architecture and a set of fully tested components to build scalable BI solutions.

On top of this foundation, iSoftStone’s BI team provides customization services, vertical knowledge and technical expertise to implement the solution based on customer’s specific needs and industries. This software plus services approach has been designed to maximize client satisfaction and provide a quick and efficiently implementable solution.

Enhancements to ETL Development & Processing

<table>
<thead>
<tr>
<th>Features</th>
<th>Other ETL Tools</th>
<th>RapidStart BI Solution Accelerator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamic Package</td>
<td>Do not offer this as an out of box solution.</td>
<td>RapidStart BI Solution Accelerator is designed to dynamically generate ETL packages based on metadata, reducing significant development efforts and providing increased maintenance efficiency. Business rules are centralized and isolated from packages and metadata allowing easy troubleshooting and manageability.</td>
</tr>
<tr>
<td>creation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full/Incremental</td>
<td>Require extensive development time and schema</td>
<td>The solution is specifically designed to handle this scenario as an out–of-the-box and easy to implement solution.</td>
</tr>
<tr>
<td>Loads</td>
<td>modifications to incorporate.</td>
<td></td>
</tr>
<tr>
<td>Metadata Management</td>
<td>Building metadata management in a project requires</td>
<td>The BI Solution Accelerator framework provides metadata management structures and components. Metadata for the entire ETL process is stored in centralized ETL Controller database tables for easy retrieval.</td>
</tr>
<tr>
<td></td>
<td>significant development efforts to implement the</td>
<td></td>
</tr>
<tr>
<td></td>
<td>solution.</td>
<td></td>
</tr>
<tr>
<td>Features</td>
<td>Other ETL Tools</td>
<td>RapidStart BI Solution Accelerator</td>
</tr>
<tr>
<td>------------------</td>
<td>--------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Logging</td>
<td>Most ETL implementations do not support centralized logging. No centralized logging of distributed ETL without extensive customization.</td>
<td>BI Solution Accelerator offers a centralized location to view log details of ETL processes running on multiple servers. SQL Server Broker Communication Framework logs run time statuses in ETL Controller database tables. All errors from all servers are available in one place.</td>
</tr>
<tr>
<td>Auditing/Data Validation</td>
<td>Require coding, not offered as an out of box solution.</td>
<td>Out of box auditing through an audit ID at the data level to track changes.</td>
</tr>
<tr>
<td>Data Lineage</td>
<td>Most tools provide only part of this functionality.</td>
<td>Tracing data transformations through an ETL pipeline is challenging. RapidStart BI Solution Accelerator offers data lineage transparency providing an end to end picture of data transformations in homogeneous and heterogeneous ETL environments.</td>
</tr>
<tr>
<td>Monitoring</td>
<td>SQL job execution visibility is limited to a single server. Broader visibility may require a customized Operations Management solution.</td>
<td>An ETL Monitor that is part of the framework allows the ability to see the job execution status as it is happening on single server or multiple servers.</td>
</tr>
<tr>
<td>Reusability</td>
<td>Need to create data flow for each individual step which makes it difficult to reuse components for different data sources without making changes.</td>
<td>Reusable code is used for source and destination definition.</td>
</tr>
</tbody>
</table>
Components Overview

Diagram showing the components of a data process flow, including Batch Script, Stored Proc, ETL Controller DB, Dynamic Packages, SQL Server, Dynamic Packages, Excel, Flatfile, SharePoint, Dynamic Packages, Staging DB, Data Mart, Cube, and ETL Monitor.
Dynamic Package Generation

The ETL Workflow is a console application (DeltaExtractor.exe) which orchestrates the entire data flow. SSIS packages are dynamically generated based on the batch and step information stored in the ETL Controller database. As soon as the job execution starts, the workflow automatically recognizes the environment (DEV, SIT, UAT or Production) in which the job is running. The workflow processes the data based on the configuration of global parameters, batch metadata and step metadata. The framework allows both parallel and sequential execution to maximize efficiency. Any data source type that is currently supported by SSIS is supported by RapidStart BI Solution Accelerator. In addition, the framework also manages execution of command shell operations, PowerShell scripts, and calls to external tools (i.e. BCP, SqlCmd, etc.).

Metadata Management

RapidStart BI Solution Accelerator offers advanced metadata management. The Batch workflow is a script with a series of steps which defines the specific ETL process. For example, one batch workflow container could be responsible for loading all dimensions in a data warehouse, while a single step in the batch can correspond to loading a dimension table. Each metadata table in the ETL Controller database contains details relevant to a batch at the global level and at the more granular step level. The workflow batch metadata provides error handling, efficient workflow execution capabilities, workflow dependencies and constraints. The workflow uses step metadata for efficiently executing steps in parallel/sequential, to identify stored procedure specific to the step and to identify/handle dependencies/constraints for each step. The metadata tables also allow the ability to configure full or delta data loads and also store runtime statuses. All batch, steps, and runtime metadata exist in one single location - the ETL Controller database.

Distributed Workflow Processing

ETL processes can be dynamically executed on different servers across the network. This capability allows for efficient use of commodity servers to expedite distributed processing of ETL workflow pipelines. The ETL Controller database is a centralized location to monitor workflow execution, view run time status, and view source to destination data transformations as it is happening on these distributed servers. The Controller intelligently assigns workflows processes out to slave servers and uses internal logic to determine how batch steps can be broken out into atomic work units. This unique capability vastly increases the power of SSIS to process large amounts of data into a database quickly and efficiently.
Error Handling
The framework uses Service Broker Communication framework to ensure the steps communicate their status back to the ETL Controller to mark their status or report error. The Controller will wait for all active threads to communicate their exit status. Default behavior is to abort workflow execution on any step failure. Workflow can be restarted and will pick up from the last failed or unfinished step. Before the workflow exits, it will try to execute On Success/On Failure Batch level processing, if defined.

Monitoring
The ETL Monitor is a user interface where users can visually monitor the progress of a workflow. The monitor displays status at batch level and step level. Detailed error logging allow for effective troubleshooting and timely resolution. The information displayed in the ETL Monitor is also available in the logging tables within the ETL Controller database, thereby making it available for reporting purposes. In addition, the ETL Monitor tool allows the ability to view statuses of execution steps running on multiple servers simultaneously. The combination of ETL Controller runtime tables, log table and ETL Monitor can help troubleshoot not only workflow and system issues but allows the ability to troubleshoot discrepancies at the data level with each job run. Having the ability to see all the information in one single location empowers analysts and support team to troubleshoot issues much faster.

Benefits
RapidStart BI Solution Accelerator’s key benefits are:

- **Rapid Development and Configuration**: iSoftStone delivers higher efficiency to its customers by reducing development time when working with SSIS and providing rapid initial standup, configuration and faster data load time.
  - ETL processes are defined and controlled via metadata stored in SQL Server (ETL Controller database).
  - Transparent dynamic package creation (with the option to save the package or delete after use).
  - Reusability of data load scripts once built.
  - Default configurations are preset and can easily be modified.
- **Scalability**: RapidStart BI Solution Accelerator improves SQL Server Integration Services scalability by enabling the distribution of ETL workflow processes across multiple commodity servers.
- **Rapid parallel data load**: The framework has the ability to create and process multiple packages in parallel and can be adjusted based on technical requirements, data volumes and available hardware resources.

- **Reduced Maintenance cost**: By dynamically creating packages, RapidStart BI Solution Accelerator reduces the number of workflows created and streamlines the maintenance needed for ETL pipelines.
  - Pipelines designed using SQL scripts and stored procedures.
  - SQL Server 2008 and SSIS are the only requirements.
  - Code management efficiency
  - Each environment (DEV, SIT, UAT, and PROD) details are setup once upon initial setup and can be easily modified if changes are required or additions are needed.
  - The Batch workflow can determine environment specifics relevant to local server, based on metadata stored within the controller database.

- **Increased audit metadata visibility**
  - Complete control over amount and type of metadata that is captured throughout the ETL process. Very useful for monitoring and audit purposes.
  - Metadata from each step in the ETL process is saved to tables and is easily accessible.

- **Runtime debugging/traceability**
  - The runtime monitor can be used to visually monitor ETL (Batch and Step) progress. It can also be used to pinpoint the cause of failures for quick resolution.
  - Configurable runtime monitor history retention.
Conclusion

RapidStart BI Solution Accelerator increases efficiency in building data warehouse solutions in small, medium, large and enterprise scale organizations facilitating a more efficient development process in conjunction with SSIS. The benefits of data lineage, monitoring, dynamic package generation, distributed data processing and metadata management have been significant in improving development speed and resources utilization. This in turn leads to executives and managers receiving business critical data faster and at a far lower cost to the organization. With its software plus services approach, iSoftStone’s BI team maximizes customer satisfaction and provides a quick and efficiently implementable business intelligence solution.

For questions about RapidStart BI Solution Accelerator please contact:
USNAERapidStart@iSoftStone.com