Governet

ASSIST: NEXT GENERATION

DATA WAREHOUSE AND DATA MART ARCHITECTURAL OVERVIEW

Version 1.0

10/7/2014

VERSION HISTORY

Version #	Implemented By	Revision Date	Approved By	Approval Date	Reason
1.0	***Toni Becvinovski***	***10/7/2014***	***Toni***	***10/7/2014***	***NA***

Table of Contents

1. Overview

1.1. Background

This documentation is intended for Governet or ASSIST technical team members who would like an understanding of the algorithms, assumptions, and vision for the Assist Data Warehouse (DW). It is assumed that the audience has an understanding of SQL Server development terms and concepts. This document contains generic descriptions of the algorithms. For actual implementation details of the concepts mentioned here, please refer to the source code. Where appropriate, actual object names are included to make that research easier.

1.2. Format

Except where noted, the information in this document represents the current architecture. When the target architecture has not been implemented yet, both the current and target architecture are explained.

1.3. ETL

A SQL Server job on the warehouse server calls a process on the Data Warehouse (DW) that truncates and rebuilds all the DW tables which contain ASSIST data. All of the ETL logic exists in the DW database so the data is pulled from the DW rather than pushed from the OLTP.

1.4. Latency

The ETL process runs nightly. By choosing a 24 hour update schedule, the code can be drastically simplified over an algorithm that would need to incrementally rebuild the data warehouse tables as it changed on the OLTP database.

1.5. Hardware/Environment

Current

The OLTP and DW databases are both hosted on the same SQL instance for each ASSIST environment.

Vision

- Host the DW on hardware that is independent from the OLTP hardware.
- Use a TBD technology, probably SQL Server Transactional Replication or Linked Servers to make a copy of the OLTP data available on the DW server for use in the ETL process

2. Data Warehouse Structure

2.1. Rules

The DW should not be the system of record for any information. All data should be derived from other sources. We should be able to drop a DW database and recreate it from source control without losing any data.

2.2. Goals

There are two main goals that drive the development of the data warehouse:

- Performance- That is to say, the time between a report request from a user and report delivery in the browser should be measured in seconds and not minutes. There are no specific success metrics that would apply to every report, but in general, an effort should be made to improve the speed of any report that takes more than a few seconds to render.
- Ease of BI implementation - The DW should make the creation of new reports and the satisfaction of ad hoc queries as easy as reasonably possible.

2.3. Tables

To support the goal of easy BI implementation, the normalized OLTP schema will be denormalized into a simpler Object based schema. This schema can serve as the basis for both transactional reports, e.g. A list of Courses, or for Dimension and Fact tables that might support more OLAP style analysis in a specialized Data Mart. Since the DW is not the system of record, data duplication in multiple Data Marts is an acceptable way to address either DW goals (performance or BI implementation).

2.4. Challenges

Almost every “object” in ASSIST’s OLTP database has an independent timeframe. Using Courses as an example, a Course would contain a Start date and End date for every version of that course. Additionally, a course has a Prefix that has its own Start and End date. Other things might relate to Courses through Prefixes that would also have their own Start and End date. Finding the Course, Prefix, etc. that was active on a specific date is pretty easy, but merging all those data sets together into one with a single set of Start and End dates is complex.

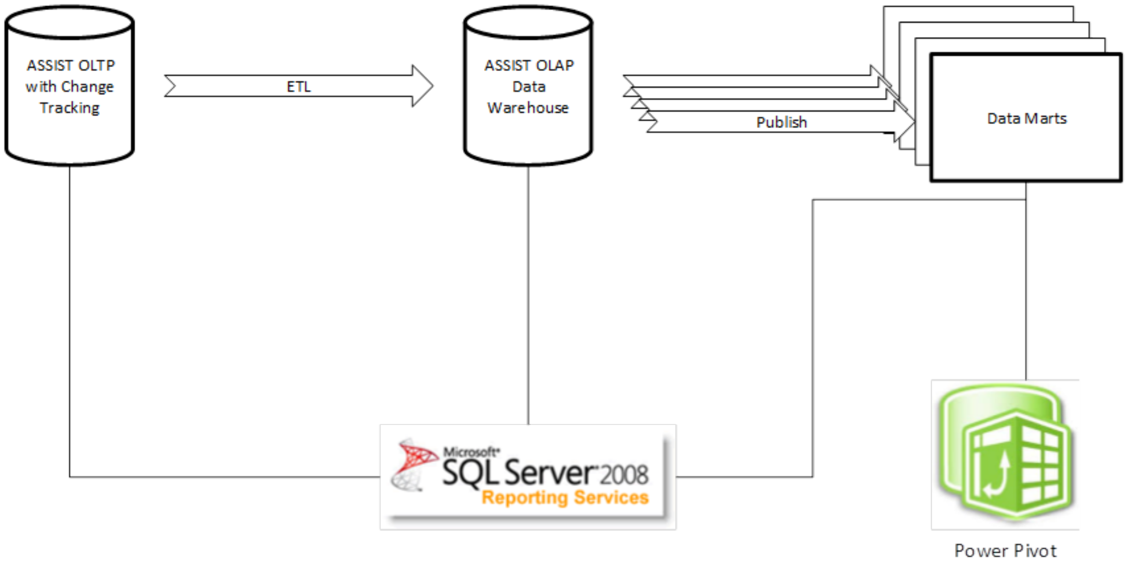
Since the algorithm to perform that merge is complex and prone to errors, a method was needed to leverage the algorithm multiple times to combine many datasets without duplicating the code. That would allow the algorithm to be improved or fixed in one place without taking the chance of introducing human errors in the maintenance or having inconsistent business rules in different merges.

To accomplish this, the algorithm was implemented as a stored procedure that merges two views. The views act as a C# interface in that they define a contract between the procedure and the calling process. It also allows for the easy testing of the merge algorithm by allowing the creation of test input data that can be fed into the procedure to see if the expected result was achieved. The procedure is called build.DataSetMerge.

2.5. Schemas

- Build – objects that are involved in the ETL process
- Test – objects that are involved in testing either ETL functionality or data integrity
- Dbo – objects related to the actual ASSIST database

Diagram of Data Warehouse and Power Point Tool


ASSIST OLTP with Change Tracking
ETL
ASSIST OLAP Data Warehouse
Publish
Data Marts
Microsoft SQL Server 2008 Reporting Services
Power Pivot

3. SDLC

3.1. Development

The development of the DW code is done using Visual Studio database projects. The project can be opened using the free version of Visual Studio, specifically: Visual Studio Express with Update 3 for Windows Desktop.

3.2. Deployment

Current - The project will be manually deployed on demand using Visual Studio.

Vision- The project will be deployed automatically using the same rules as other Governet database deployments.

4. Ad Hoc Support

Vision

The DW will support ad hoc reporting by exposing a subset of the DW to users as one or more PowerPivot Data Marts (DM). The DMs will be designed around a specific ad hoc use case or set of use cases. It will be optimized for PowerPivot by changing of data types and column names to include better descriptions. Typically, it would only contain one “table” (which will frequently be implemented as a view) that contains the data necessary to satisfy the use cases that have been targeted.

Delivery of the Excel sheet containing the DM data in a PowerPivot format will be done through the application as an on demand download. It should be noted that after the data has been downloaded by an authorized user, that user has full control over the data and its distribution. Therefore, access to the ad hoc reporting spreadsheets should only be granted to a small number of users.