Cleaning, Validating and Enhancing the SQL Server Data Warehouse Contact Dimension

Ray Barley

Business Intelligence Architect

Agenda

- Overview
- Melissa SSIS Data Components
- Contact Dimension Solution and Demo
- Questions

Overview

Problem Statement

- We get Contact data from a variety of sources; e.g. trade shows, web sites, etc.
- Contact data may not be complete and in a standardized format
- We need to clean, validate and enhance (aka Transform) the Contact data we receive and update the Contact dimension in a SQL Server data warehouse
- How do we leverage SSIS to do this?

Overview
Using SSIS

- SSIS provides built-in components to perform Extract, Transform and Load (ETL) operations
- SSIS components generally provide generic capabilities; we need domain-specific capabilities; i.e. cleaning, validating and enhancing Contact data
- SSIS provides scripting components for domain-specific logic (you write your own .NET code)

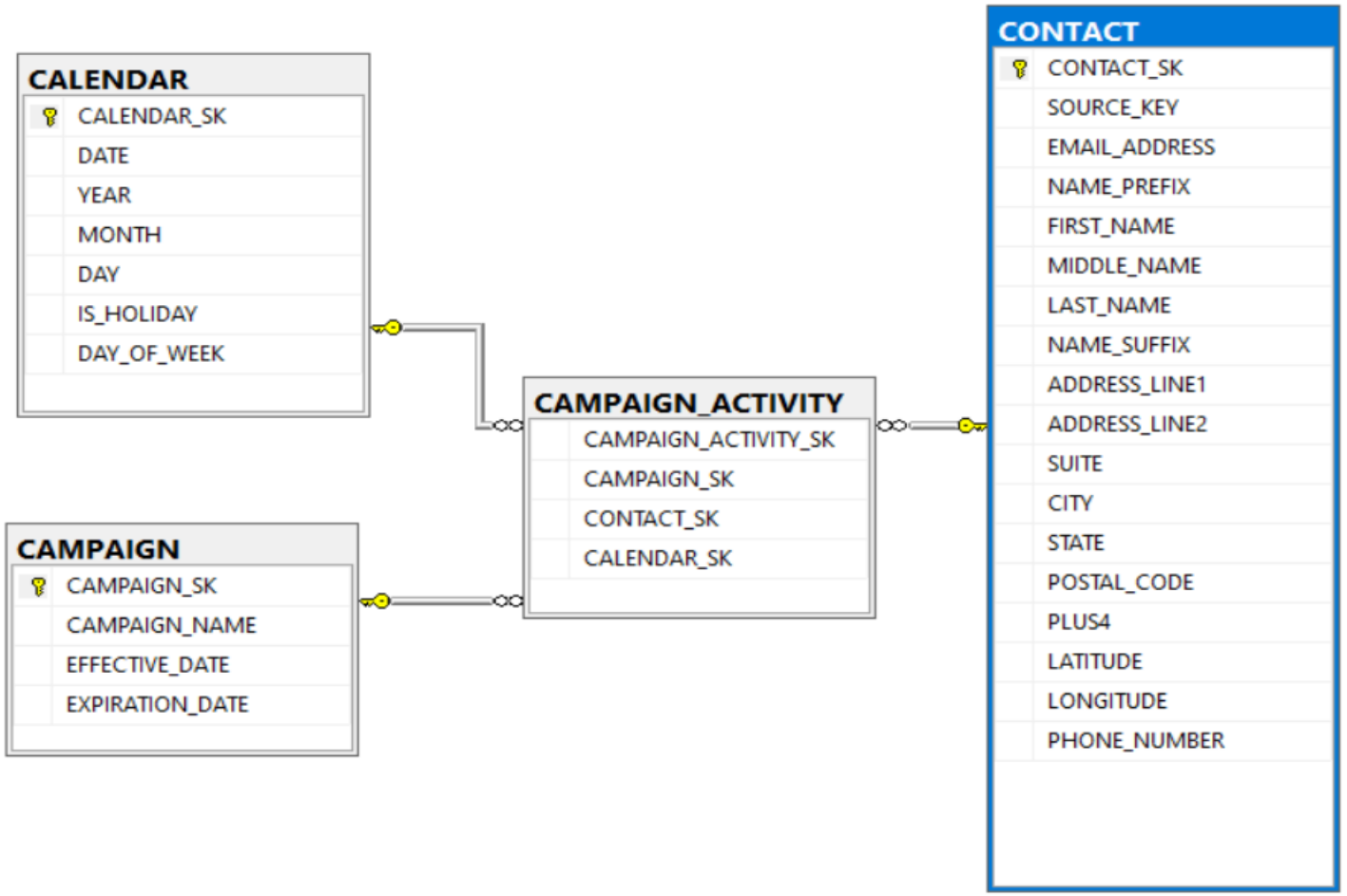
Overview

Melissa SSIS Components

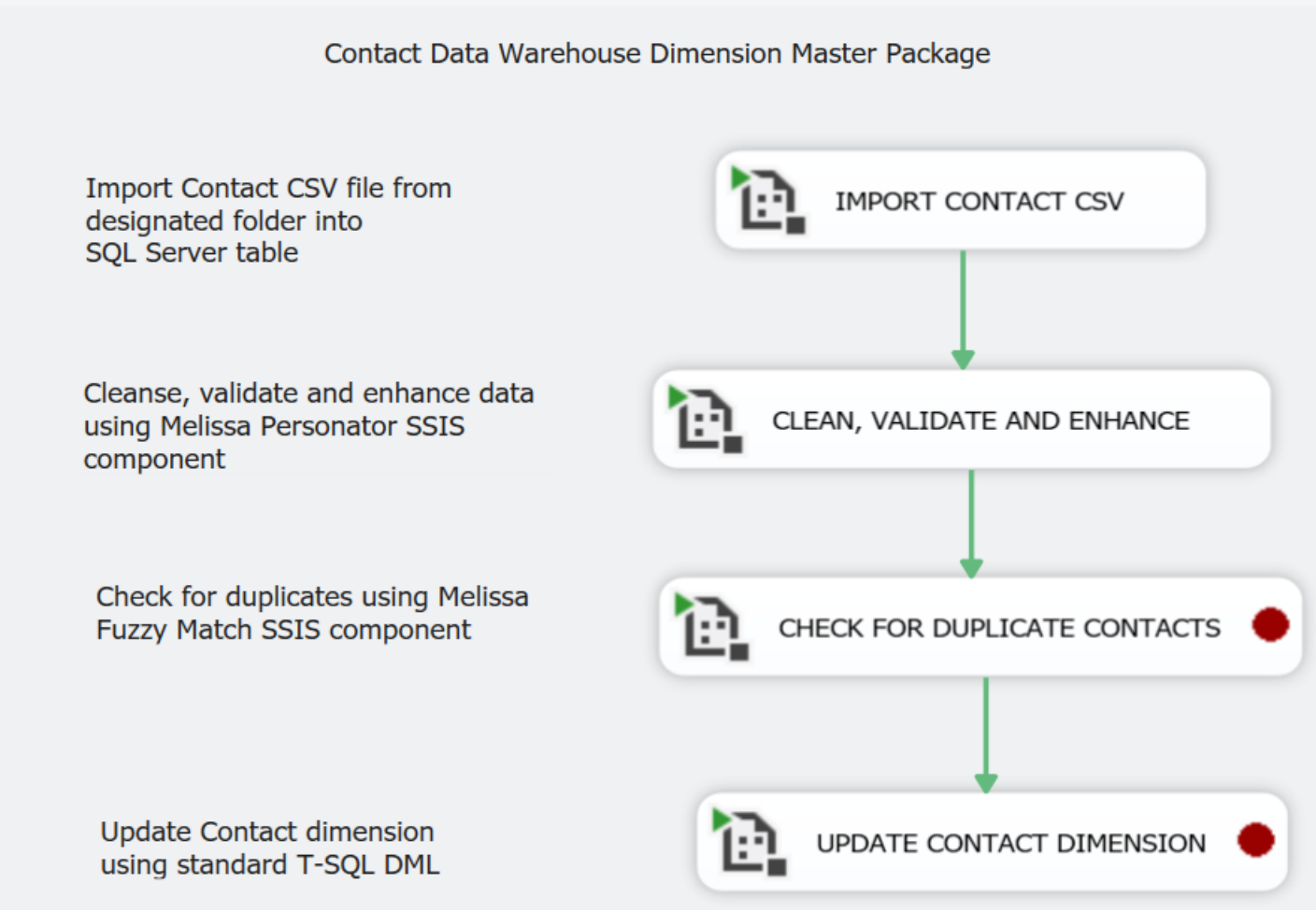
- SSIS components with domain-specific knowledge of Contact data cleansing, validating and enhancing
- Plug and play; add to SSIS Data Flow; no coding required!
- Wealth of capabilities including:
  - Properly parse names, addresses, email addresses, and phone numbers
  - Validate addresses, email addresses and phone numbers
  - Add enhanced contact information including latitude/longitude coordinates and demographic data

Overview

Simplified Contact Dimension



Overview
What will we learn today?



Melissa SSIS Components

- Overview
- Profiler
- Personator
- Fuzzy Match

Melissa SSIS Components
Overview

- SSIS Data Flow components; work with any source or destination
- Robust component configuration; save configuration for reuse
- Built-in filtering to determine good or bad records; can be customized
- Component-specific result codes provide detailed information on success or failure

Melissa SSIS Components
Profiler

- Perform deep analysis of data on several levels:
  - Formatting – does the data ‘look’ like what is expected
  - Content – uses reference data to determine if the data is consistent with what is expected
- Add into the SSIS Data Flow
- Can analyze data from any source supported by SSIS and 3rd party vendors
- Save output to any SSIS data destination

Melissa SSIS Components
Personator

- All-in-one contact verification and appending
- Works on names, addresses, phone numbers and email addresses
- Parses, corrects, and adds derived data
- Verifies name corresponds to address, email and phone
- Appends missing name and company name, phone, and email
- Provides current addresses for people and companies that have moved with 20+ years of history

Melissa SSIS Components

Fuzzy Match

- Identify duplicate records from multiple sources
- User-defined data can be matched and duplicates eliminated
- Large toolbox of state-of-the-art fuzzy matching algorithms
- Granular control on match thresholds and fine tuning of algorithms
- Based on percentage score between records, records are directed to Matches, Possible matches and Non-Matches

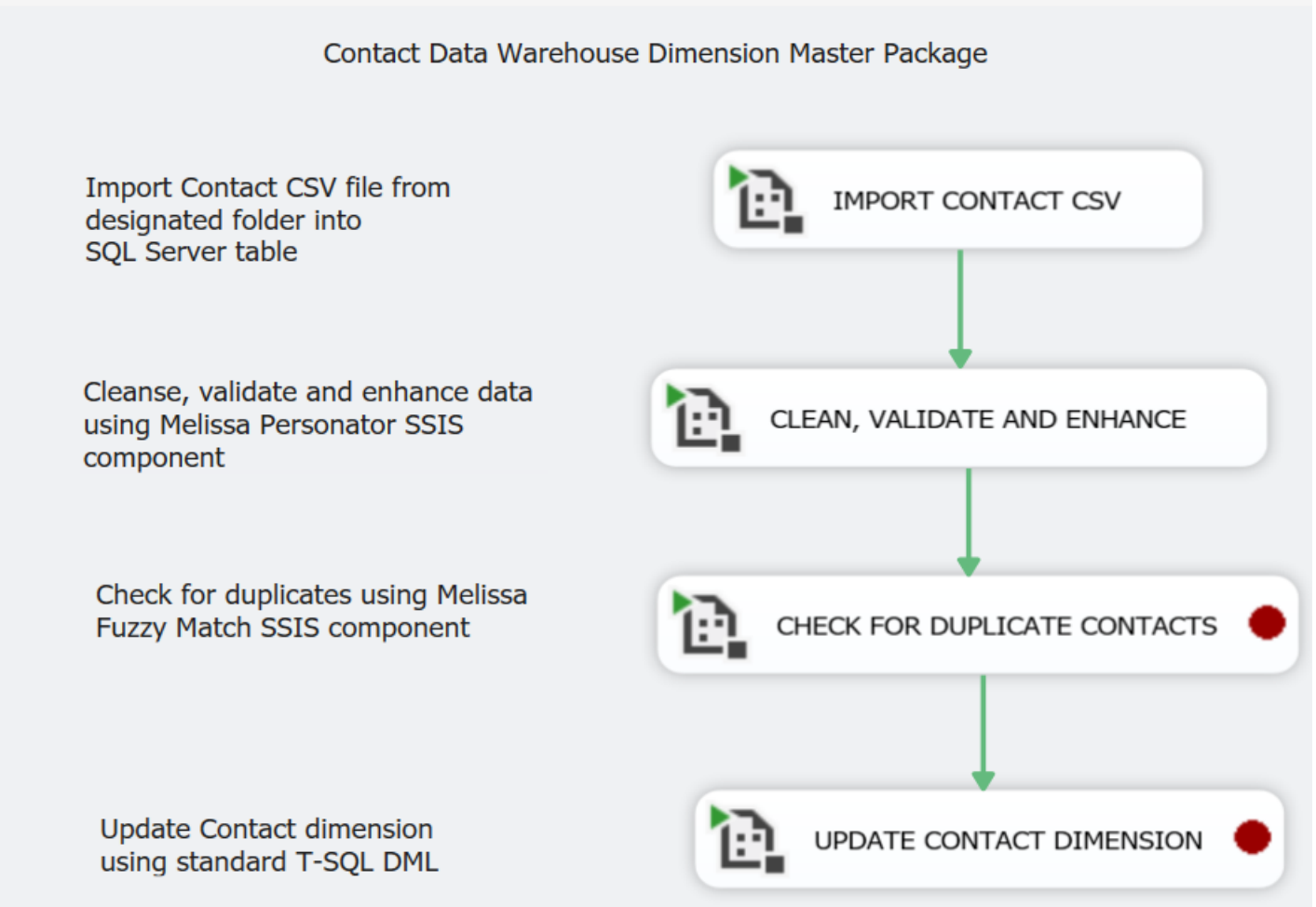
MSSQLTips

Contact Dimension Solution

- High-level solution
- Demo

Contact Dimension Solution

High-level solution



Contact Dimension Solution
Demo

SSIS Solution Demo

Questions

References

- Melissa SSIS Data Quality Components

http://wiki.melissadata.com/index.php?title=SSIS%3AData_Quality_Components

https://www.melissa.com/resources/data-sheets/pdf/dqt-ssis-brochure.pdf

http://wiki.melissadata.com/index.php?title=Result_Codes