

Grumpy Old IT Guy SQL Server Matching Architecture Framework (GITG-SMAF™)

A Practical Framework for Enterprise Record Matching Using Microsoft SQL Server

Version 1.0

**Copyright © 2026 Grumpy Old IT Guy Publishers /
Angelo R. Bobak**

Executive Summary

Organizations rarely maintain a single, perfectly clean source of customer, supplier, product, employee, or asset information. Instead, information is collected from numerous operational systems, mergers, acquisitions, cloud applications, spreadsheets, and external providers. As a result, duplicate, incomplete, inconsistent, and conflicting records become inevitable.

Traditional exact matching techniques quickly fail when confronted with real-world data containing typographical errors, abbreviations, missing values, transposed characters, nicknames, formatting differences, or changing business information. Enterprise organizations therefore require sophisticated matching architectures capable of determining whether two imperfect records actually represent the same real-world entity.

The **Grumpy Old IT Guy SQL Server Matching Architecture Framework (GITG-SMAF™)** provides a structured methodology for designing, implementing, and governing enterprise record matching solutions using native Microsoft SQL Server technologies. Rather than relying on a single matching algorithm, the framework combines data profiling, cleansing, standardization, candidate generation, multiple similarity algorithms, configurable business rules, metadata, trusted reference data, and human stewardship to maximize matching accuracy while maintaining transparency and governance.

The result is a repeatable architecture capable of producing trusted **Golden Records** that become authoritative business entities throughout the enterprise.

1. Introduction

Record matching is one of the most difficult problems in data management.

Organizations often assume duplicate records can be identified by comparing names or identification numbers. In reality, two records describing the same customer may contain:

- Different spellings
 - Missing addresses
 - Nicknames
 - Alternate company names
 - Multiple phone numbers
 - Outdated information
 - Formatting inconsistencies
 - Data entry errors
 - OCR mistakes

- Legacy system abbreviations

Because no single algorithm can accurately resolve every situation, successful matching solutions combine multiple techniques into an integrated architecture.

The GITG-SMAF™ provides that architecture.

2. Framework Objectives

The framework is designed to:

- Improve enterprise data quality
- Reduce duplicate records
- Increase customer accuracy
- Support Master Data Management
- Improve regulatory compliance
- Reduce operational costs
- Increase confidence in enterprise reporting
- Enable trusted AI training datasets
- Produce sustainable matching processes
- Support long-term governance

3. Core Principles

The Grumpy Old IT Guy SQL Server Matching Architecture Framework (GITG-SMAF™) is built on eight foundational principles that guide the design of accurate, scalable, and trustworthy enterprise matching solutions.

Principle 1

Profile before matching.

Never attempt matching before understanding data quality.

Principle 2

Clean before comparing.

Standardization dramatically improves matching accuracy.

Principle 3

Reduce the search space.

Blocking and candidate generation dramatically improve performance.

Principle 4

Multiple algorithms outperform one algorithm.

Different fields require different similarity techniques.

Principle 5

Business rules matter.

Matching is a business decision—not merely a mathematical calculation.

Principle 6

Metadata provides context.

Attribute quality influences confidence.

Principle 7

Humans remain part of the process.

Not every decision should be automated.

Principle 8

Golden Records are governed assets.

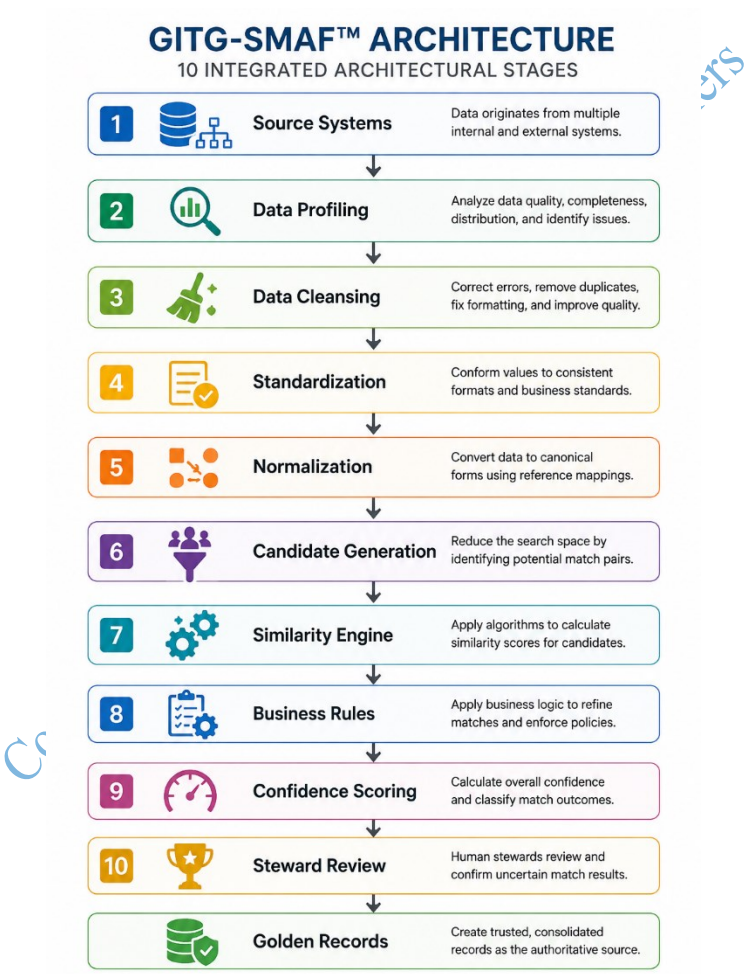
Matching does not end with identifying duplicates.

You have to fix them!

The framework emphasizes understanding and improving data quality before matching, reducing unnecessary comparisons through intelligent candidate generation, combining multiple similarity algorithms with business rules and metadata, and incorporating human stewardship where automated decisions are insufficient. Ultimately, the objective extends beyond simply identifying duplicate records to creating governed Golden Records that serve as trusted, authoritative business assets across the enterprise.

4. GITG-SMAF™ Architecture

The framework consists of ten integrated architectural stages.



Source Systems

Enterprise data originates from multiple internal and external systems, including CRM, ERP, finance, e-commerce, spreadsheets, cloud applications, and third-party data providers. These sources often contain inconsistent, incomplete, or duplicate representations of the same business entity.

Data Profiling

Data profiling examines the incoming data to understand its overall quality, structure, completeness, uniqueness, and consistency. The results identify potential issues that must be addressed before accurate record matching can occur.

Data Cleansing

Data cleansing corrects obvious errors such as misspellings, invalid values, duplicate whitespace, formatting inconsistencies, and incomplete records. Improving data quality before matching significantly increases the accuracy of subsequent comparisons.

Standardization

Standardization converts similar business values into a common, consistent format using predefined business rules and reference tables. Examples include standardizing street abbreviations, company suffixes, state names, telephone formats, and date representations.

Normalization

Normalization transforms data into canonical forms that eliminate unnecessary variations while preserving business meaning. Common techniques include removing

punctuation, converting text to a consistent case, expanding abbreviations, and mapping nicknames or alternate spellings to standardized values.

Candidate Generation

Rather than comparing every record with every other record, candidate generation intelligently narrows the search space by grouping records that are likely to match. Blocking keys, phonetic encoding, geographic regions, and other filtering techniques dramatically improve performance while maintaining high matching accuracy.

Similarity Engine

The similarity engine compares candidate records using one or more matching algorithms designed to measure how closely individual attributes resemble one another. Depending on the data being evaluated, techniques such as edit distance, phonetic comparison, token analysis, n-gram matching, or native SQL Server fuzzy matching functions may be applied.

Business Rules

Business rules incorporate organizational knowledge into the matching process by validating or rejecting potential matches based on business-specific conditions. These rules ensure that mathematical similarity is balanced with real-world business logic, regulatory requirements, and operational policies.

Confidence Scoring

The results from similarity algorithms and business rules are combined into an overall confidence score representing the likelihood that two records describe the same real-world entity. Configurable score thresholds classify each comparison as an automatic match, a candidate for steward review, or a confirmed non-match.

Steward Review

Records with uncertain confidence scores are evaluated by data stewards who use business expertise to make the final matching decision. Steward actions also provide valuable feedback that can be used to refine matching algorithms, business rules, and scoring thresholds over time.

Golden Records

Confirmed matches are consolidated into trusted Golden Records that represent the authoritative version of each business entity. These records have become the enterprise's single source of truth, providing consistent, reliable data for operational systems, reporting, analytics, artificial intelligence, and data governance initiatives.

5. Framework Lifecycle

The framework follows a continuous lifecycle.



Phase 1 – Data Profiling

Objectives include:

- Identify missing values
- Discover duplicates
- Measure completeness
- Detect inconsistent formatting
- Analyze attribute distributions
- Identify invalid values
- Assess data quality

Typical SQL Server technologies include:

- T-SQL profiling queries
- SQL Server Reporting Services
- Data Quality Services
- SQL Server Integration Services

Deliverables:

- Data Profile Report

- Quality Scorecards
- Attribute Analysis

Phase 2 – Data Cleansing

The objective is to correct data before matching.

Activities include:

- Correct invalid values
- Repair formatting
- Remove illegal characters
- Resolve abbreviations
- Validate postal information
- Correct capitalization
- Remove duplicate whitespace

Deliverables:

- Cleansed dataset
- Data quality metrics
- Cleansing audit logs

Phase 3 – Standardization

Standardization ensures consistent representation.

Examples include:

Street
ST
St.
Street.

↓

STREET

Examples:

International Business Machines

↓

IBM

Deliverables:

- Standardized attributes
- Business dictionaries
- Domain rules

Phase 4 – Normalization

Normalization converts multiple representations into a common format.

Examples include:

Phone numbers

(914) 555-1234
914.555.1234
9145551234

↓

9145551234

Dates

07/04/2026

↓

2026-07-04

Names

ROBERT

↓

BOB

Using nickname reference tables.

Deliverables:

- Normalized dataset
- Lookup tables
- Reference mappings

Phase 5 – Candidate Generation

Comparing every record against every other record is computationally expensive.

Instead, candidate generation limits comparisons through:

- Blocking keys
- Soundex groups
- ZIP code partitions
- Initial character groups
- Date of birth partitions
- Geographic regions
- Business classifications

Benefits include:

- Faster execution
- Reduced CPU usage
- Better scalability

Phase 6 – Similarity Analysis

Candidate records undergo similarity analysis using multiple algorithms.

Typical algorithms include:

- Levenshtein Distance
- Damerau-Levenshtein
- Jaro
- Jaro-Winkler
- N-Gram Similarity
- Edit Distance
- Phonetic Matching
- Token Matching
- SQL Server 2025 native fuzzy matching functions

Each attribute may use a different algorithm.

Phase 7 – Business Rule Evaluation

Mathematical similarity alone rarely produces reliable matches.

Business rules refine the decision.

Examples include:

IF

Last Name > 95%

AND

Birth Date Exact

THEN

Increase confidence

IF

Social Security Numbers differ

THEN

Reject match

Business rules reflect organizational policies.

Phase 8 – Confidence Scoring

Each comparison produces an overall confidence score.

Example

Name Similarity	35%
Address Similarity	25%
Phone Similarity	15%
Email Similarity	15%
Business Rules	10%

Overall Score	92%

Thresholds classify records.

For example:

95–100%

Automatic Match

85–94%

Steward Review

Below 85%

Non-Match

Thresholds should be configurable to reflect business risk and tolerance.

Phase 9 – Steward Review

Not every comparison should be automated.

Data stewards review uncertain candidates.

Responsibilities include:

- Accept match
- Reject match
- Merge records
- Document rationale
- Improve business rules
- Refine thresholds

This feedback becomes part of continuous improvement.

Phase 10 – Golden Record Creation

Confirmed matches are merged into trusted business entities.

Golden Records become:

- Customer master
- Product master
- Supplier master
- Employee master
- Asset master

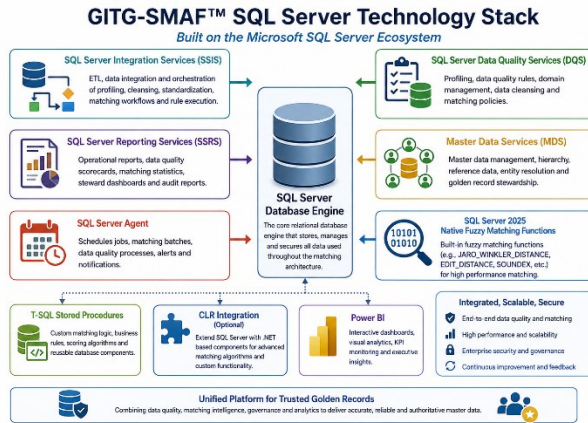
Survivorship rules determine which attribute values become authoritative.

Examples include:

- Most recent address
- Highest quality phone number
- Verified email
- Trusted source priority
- Steward-approved values

6. SQL Server Technology Stack

The framework leverages the Microsoft SQL Server ecosystem.



Typical components include:

- SQL Server Database Engine
- SQL Server Integration Services (SSIS)
- SQL Server Data Quality Services (DQS)
- SQL Server Reporting Services (SSRS)
- Master Data Services (where available)
- SQL Server Agent
- SQL Server 2025 native fuzzy matching functions
- T-SQL stored procedures
- CLR integration (optional)
- Power BI

7. Governance Integration

Matching should operate within a governed environment.

Key governance components include:

- Data Owners
- Data Stewards
- Metadata Repository
- Business Glossary
- Reference Data Management
- Audit Logging
- Version Control
- Approval Workflows
- Policy Management
- Compliance Monitoring

8. Deliverables

A complete GITG-SMAF™ implementation typically produces:

- Enterprise Matching Architecture
- Data Profiling Report
- Cleansing Standards
- Standardization Rules
- Normalization Dictionaries
- Blocking Strategy
- Matching Rule Catalog
- Similarity Algorithm Matrix
- Business Rule Repository
- Confidence Score Matrix
- Steward Review Workflow
- Golden Record Model
- Survivorship Rules
- Metadata Repository
- Governance Procedures
- Operational Dashboards
- Matching Performance Reports

9. Benefits

Organizations implementing the GITG-SMAF™ framework can expect:

- Higher matching accuracy
- Fewer duplicate records
- Improved customer 360° views
- More trusted analytics
- Better regulatory compliance
- Reduced manual reconciliation
- Increased operational efficiency
- Higher quality AI training datasets
- Stronger data governance
- Enterprise-wide consistency

10. Implementation Roadmap

Phase 1 – Assessment

- Profile existing data sources
- Assess data quality
- Identify duplicate patterns

Phase 2 – Foundation

- Define matching policies
- Create reference data
- Establish business rules

Phase 3 – Solution Development

- Implement cleansing
- Standardize and normalize data
- Develop candidate generation

- Configure similarity algorithms

Phase 4 – Validation

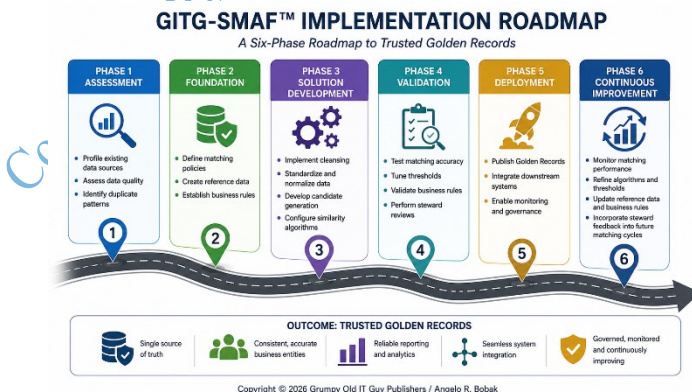
- Test matching accuracy
- Tune thresholds
- Validate business rules
- Perform steward reviews

Phase 5 – Deployment

- Publish Golden Records
- Integrate downstream systems
- Enable monitoring and governance

Phase 6 – Continuous Improvement

- Monitor matching performance
- Refine algorithms and thresholds
- Update reference data and business rules
- Incorporate steward feedback into future matching cycles



Conclusion

The **Grumpy Old IT Guy SQL Server Matching Architecture Framework (GITG-SMAF™)** recognizes that successful enterprise record matching is not the result of a single algorithm, but of a disciplined architecture that integrates data quality, similarity analysis, business knowledge, and governance into a cohesive process. By progressing through profiling, cleansing, standardization, normalization, intelligent candidate generation, configurable similarity scoring, business rule evaluation, stewardship, and Golden Record creation, organizations can transform fragmented and inconsistent data into trusted, authoritative business entities.

The framework is intentionally technology-focused while remaining methodology-driven. It leverages native Microsoft SQL Server capabilities alongside proven data management practices to create a scalable, explainable, and auditable matching solution. Whether applied to customer, supplier, product, employee, or other master data domains, GITG-SMAF™ provides a repeatable blueprint for building high-confidence matching systems that support operational excellence, regulatory compliance, analytics, artificial intelligence, and enterprise data governance.

One of the central principles of the framework reflects a lesson learned from countless real-world implementations:

"Matching is not about finding identical records. Matching is about discovering the truth hidden behind imperfect data."

This philosophy captures the essence of GITG-SMAF™—combining technology, business expertise, and governance

to identify the most accurate representation of every real-world entity while continuously improving the quality and trustworthiness of enterprise data.

© 2026 Angelo R. Bobak
All rights reserved.

First Edition: February 2026

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopying, recording, or otherwise—without the prior written permission of the publisher.

Grumpy Old IT Guy Publishers
Westchester County, NY, USA
www.grumpyolditguy.com