Advances in IT and Digital Security, Volume 1, Issue 8, pp. 1-35, July 2026.

Advances in IT and Digital Security
ISSN 3143-4266
DAC Insight Publishers
https://journals.dacinsightpublishers.com/AIDY

Validation, Deduplication, and Reconciliation Controls in Large-Scale CRM Data Migration: A Critical Review

Abdullahi Olalekan Abdulkareem[1*]
[1]Independent Researcher, USA
*Corresponding author: abdullahi.abdulkareem.ola@gmail.com

ABSTRACT

Data migration of millions of customer records into one customer relationship management (CRM) platform, and the consolidation of several business units' customer records in one place, are crucial tasks of data engineering, since any data quality problem at load time will form the single customer view which sales, services and marketing will have to operate with. This critical review will bring together research on data quality dimensions and metrics, record linkage, entity resolution, data integration, and published documents on the Salesforce platform to reveal the control points which make or break a successful migration of several million records. Migration is examined here as a sequence of five gates (profiling, cleansing, entity resolution, load and reconciliation) and it appears that the literature is good on quality dimensions and matching but poor on acceptance criteria for load and cutover. There are three common pitfalls: validation focused on field level formatting and not on record-level identity; one-time setting and no re-estimation of match threshold for each business unit; and reconciliation using control totals which will not detect offsetting errors. The contribution of this paper includes gate-based validation approach, a decision rule combining the Fellegi-Sunter match weight and clerical review band, a sample size rule for record-level audit including finite population correction and practical calculations for six million records migration across five subsidiaries.

Keywords: *CRM migration; data quality; record linkage; deduplication; reconciliation; audit sampling; Salesforce; data governance*

Received 4 June 2026; Received in revised form 18 June 2026; Accepted 25 June 2026, Available online 30 June 2026.

1. INTRODUCTION

When a bank or any other multi-unit business migrates CRM data of several subsidiaries onto one platform, it buys a single customer view: one record per customer with the history of all relationships that the organization has. As Payne and Frow (2005) say, CRM is a cross-functional strategy in which information management is a core process and Chen and Popovich (2003) prove that the CRM success depends on people, process and technology, so migration is an organizational change and not merely a data load. The commercial implications of CRM implementation in United States commercial banks are proved by Krasnikov et al. (2009) who report gains in cost and profit efficiency from such implementation and the need for clean and deduplicated customer data is proven by downstream analytics like segmentation and churn prediction (Ngai *et* al., 2009).

Migration is a point where this dependence is checked. Poor data quality generates operational, tactical and strategic costs for the typical enterprise (Redman, 1998) and Haug *et al.* (2011) classify and measure these costs. The issues related to data quality are among the most studied reasons of enterprise systems implementation difficulties, as shown by several enterprise resource planning project studies (Xu *et* al., 2002; Zahari *et* al., 2015). Migration is the time when data from independently governed sources with their individual definitions, keys and quality problems come together and anything that passes the load becomes the operational truth met by staff and customers the next day.

There is no sufficient alignment between practice and literature on controlling this point. In practice, validation mostly relies on record counts and formatting check of individual fields while the literature sees data quality dimensions (Wang & Strong, 1996), record linkage (Fellegi & Sunter, 1969), data integration (Doan *et* al., 2012) as separate topics and very few sources follow the entire migration journey through profiling, cleansing, entity resolution, load and reconciliation. However, the problem is quite practical. The author was a member of a Centre of Excellence implementing Sales, Service, Marketing and Experience Cloud across five subsidiaries of Stanbic IBTC and migrating more than six million records from May 2021 to August 2023. This scale provides the practical examples used in this paper. The review raises four questions. Which quality dimensions and metrics should control migration? How should thresholds of entity resolution be set and reviewed for each business unit? What are the constraints on load of the target platform and how do they interact with validation? How large should be the record-level audit sample and what does it license?

The paper delivers a critical synthesis, a framework with five gates and exit criteria, and four worked examples of calculations: completeness metric, match-weight and threshold rule, load plan, and sample-size rule. The paper proceeds as follows: section 2 contains a description of the methodology. Section 3 contains the critical synthesis, framework, and analysis of the literature. Section 4 discusses the implications and limitations of the work, and Section 5 draws conclusions.

2. METHODS

2.1 Review design

This is a critical review. The theoretical foundations of the topic span from 1959 to 2012, the constraints of the practice context are imposed by the platform documentation that is updated in each new release, and there is no single database that references both. A critical review enables a

balanced consideration of both classical and contemporary literature. Four source classes were used: peer-reviewed literature on data quality, record linkage, entity resolution, and data integration (Batini *et* al., 2009; Christophides *et* al., 2021); classical statistics and auditing sampling theory (Cochran, 1977; Maingot & Quon, 2009); standards and legislation: ISO/IEC 25012 (ISO/IEC, 2008), Regulation (EU) 2016/679 (European Parliament & Council of the European Union, 2016), and the Nigeria Data Protection Act, 2023 (Federal Republic of Nigeria, 2023); and Salesforce documentation on bulk loading and record access (Salesforce, n.d.-a; Salesforce, n.d.-b). Vendor documentation was consulted on 22 October 2026 and the constraints specified herein correspond to the documentation available as of that date. Sources not peer reviewed were used for the sake of factual information about the platform and the legal obligations, but not for the evaluation of the effectiveness of the control.

2.2 Practice context and worked-example design

The practice context is the migration of the Stanbic IBTC described in Section 1, which involved the data validation of the records across five subsidiaries and a load of more than six million records with minimal downtimes. Migration strategy literature defines a single cutover migration and an incremental one, where some parts of the previous application are decommissioned in several stages (Brodie & Stonebraker, 1995), and the framework developed below assumes a staged approach by subsidiary for that reason. There are no organisational data provided in this paper. Probabilities in Table 2, thresholds in Table 3, and the numbers in the load plan discussion are illustrative values chosen to illustrate the calculations, unless a number is referenced from the Salesforce documentation. Results of sample-size calculation follow from the stated formulae and are not an empirical finding.

In order to keep the boundary clearly visible in this paper: all the throughput numbers, prior defect rates, audit results, and wave allocations provided in the results are illustrative values chosen to illustrate the calculations, not measurements from the practice context, and no performance claim made in this paper should be taken as the evidence of operational effectiveness. If a number is taken from the published Salesforce documentation, it is referenced at the place of usage. Section 3.15 provides a sensitivity range for the load plan assumptions, and Section 4.8 states the measurements that need to be collected during the live migration in order to validate the framework empirically.

2.3 Extended source base

Apart from the foundations named above, the review relies on four more source classes. The first is the peer-reviewed work on data profiling and automated verification: the survey of relational data profiling (Abedjan *et* al., 2015) and the declarative data quality verification system described for large datasets (Schelter *et* al., 2018). The second is the record linkage evaluation literature, including indexing and blocking surveys (Christen, 2012b; Papadakis *et* al., 2016), the guidelines on evaluating linkage quality (Harron *et* al., 2017), and applied probabilistic linkage (Sayers *et* al., 2016; Enamorado *et* al., 2019). The third is the data governance and privacy literature, including data governance frameworks (Abraham *et* al., 2019; Khatri & Brown, 2010) and privacy models for released data (Sweeney, 2002; Machanavajjhala *et* al., 2007). The fourth is the practice literature from the supplementary seed collection, which is only briefly described through its title. Existing sources not peer reviewed could not be used.

2.4 Source selection and appraisal

The review is not systematic but the sources have been located and appraised according to an explicitly stated procedure which makes reproduction of the selection process and challenge of its validity possible. Table 1 shows, for each source class, where the search was carried out, what search strings were employed and covered, and what criterion was used to select a source. There were two principles of appraisal. A source was allowed to enter the synthesis only if it deals with at least one control applied in the gate, that is some quality dimension, some metric, some matching or blocking technique, some interval estimator, or some platform limitation; and all non-peer-reviewed sources have been selected only to provide information on platforms or fulfill some legal duties, not to support the claim of a control being effective. Adjacent-domain practice sources have been kept only if there was a particular technical argument supporting them, and their validity is discussed in Section 4.7. The most flawed step in the process was that no protocol was registered and no secondary reader was involved in the selection, which is the distinction of this procedure from systematic review.

Table 1: Procedure of source selection and appraisal for the review.

Source class	Databases and sources	Search strings and coverage	Selection and appraisal
Data quality, record linkage, entity resolution, data integration (peer reviewed)	Scopus; Web of Science; ACM Digital Library; IEEE Xplore; backward and forward citation chaining from seminal works	("data quality" AND (assessment OR migration OR dimension)); ("record linkage" OR "entity resolution") AND (blocking OR deduplication OR matching); "data cleaning" OR "data profiling"; coverage 1959-2026	Peer reviewed; admitted when it supplies a dimension, metric, matching method, or evaluation result used by a gate; seminal sources retained for foundational theory
Statistical and audit sampling	Same databases; monographs (Cochran, 1977)	(audit OR stratified OR attribute) AND sampling; "confidence interval" AND (proportion OR binomial); coverage 1927-2026	Admitted when a formula or design rule used by Gates 3 or 5 is drawn from it
Standards and law	ISO/IEC catalogue; EUR-Lex; Nigeria Data Protection Commission	Instrument titles and numbers; 2008-2023, as current at retrieval	Official texts only; each obligation mapped to a gate control in Sections 3.6 and 3.13

Source class	Databases and sources	Search strings and coverage	Selection and appraisal
Platform documentation	Salesforce developer documentation (bulk loading; record access)	Limits pages: ingest allowances, file size, lock behaviour, ownership skew; versioned pages read on 22 October 2026	Used only for platform facts; every quotation scoped to the retrieval date
Practice and supplementary collection	Titles as listed in the references	No database search; items held by the authors	Non-peer-reviewed items used only for platform and legal facts; adjacent-domain items retained only where a specific technical argument is supported (Section 4.7)

3. RESULTS

3.1 Migration as a data quality problem

Wang and Strong (1996) argued that data quality is multidimensional and is determined by the fitness of data to its users, and classified dimensions into intrinsic, contextual, representational, and accessibility categories. Later ISO/IEC 25012 standard specified the model of fifteen data quality characteristics (ISO/IEC, 2008). For a migration, the lesson is that acceptance criteria should be determined by the consumers of the data: the sales team requires complete and up-to-date contact data, the services team requires accurate history of the case linked to the correct customer, and the compliance function requires defensible identification. A migration meeting a count target still can fail all of these users.

Pipino *et al.* (2002) showed how subjective assessments and objective measures can be combined, and gave simple ratio, minimum or maximum, and weighted average forms for objective metrics. The simple ratio form is the workhorse for migration gates. For a mandatory field with N_t values examined and N_d values that are missing or fail a validity rule, the quality score is given by Equation (1), and the gate can require a floor for each field and each source.

$$Q = 1 - \frac{N_d}{N_t} \qquad (1)$$

A defect taxonomy provides the team with guidance on what to test for. Kim *et al.* (2003) categorized dirty data into missing, wrong, and non-standard kinds of data depending on the stage of occurrence, and Batini *et al.* (2009) compared methodologies for assessing and improving quality. In consolidation, there are two main defects: inconsistencies between sources due to a particular fact being described in a different way by different subsidiaries, and non-uniqueness due to the same customer existing in several databases. Non-uniqueness is considered in detail in Section 3.3.

3.2 Profiling and cleansing across heterogeneous sources

Rahm and Do (2000) distinguished the problem of data cleaning into single-source and multi-source cases, and into schema-level and instance-level problems. A consolidation of five subsidiaries is a multi-source case both at the schema level and at the instance level. At the schema level, subsidiaries represent the same concept differently, one storing a single name field while another storing three fields, one using national identifier while another one is using its own customer number. At the instance level, values are conflicting, formats vary, and populations intersect. Schema mapping and transformation are studied in data integration literature (Doan *et* al., 2012), and the practical implication of this is that profiling should be carried out for each source, with a defect register agreed with each data owner prior to transformation development.

The cleansing rules must follow the principle that an action should be taken on a value only if a deterministic rule is available, such as standardizing a telephone number to the same format worldwide or a country name to a code, and quarantine the record in case a judgement is needed to repair the defect. Quarantined records will be referred back to the data owner with a classification of defect included in the record. This ensures accountability of the automatic repair process and returns the decision about ambiguous cases to a person knowledgeable in the business domain. Furthermore, it fits the wave-by-wave staged approach suggested by Brodie and Stonebraker (1995), in which each subsidiary may conduct their wave only when the defect register of this subsidiary is closed. A dedicated field must hold the legacy identifier of every source record in the target, because it ensures idempotent loads via upsert and provides the key for the reconciliation in Section 3.5 (Salesforce, n.d.-a).

Research works in the supplementary collection provide two adjacent views on the cleansing and integration processes. A review of machine learning methods for biomolecular data wrangling and curation includes learned extraction, error detection, integration, and curation workflows (Ndupu *et* al., 2026), and a data integration methodology for greenhouse gas inventory reporting covers accuracy and reliability issues (Azeez & Badmus, 2024). Both papers come from another domains and are referenced here for the sake of the common problem of detecting and correction of errors during the integration process.

3.3 Entity resolution and deduplication

Entity resolution decides which records describe the same customer. Newcombe *et al.* (1959) introduced the idea of weighting agreement and disagreement on identifying fields, and Fellegi and Sunter (1969) gave it a formal basis. For a field i with probability m_i of agreeing among true matches and probability u_i of agreeing among non-matches, the weights for agreement and disagreement are given by Equation (2), and the score for a candidate pair is the sum of the field weights, given by Equation (3).

$$w_i^+ = \log_2 \frac{m_i}{u_i}, \quad w_i^- = \log_2 \frac{1-m_i}{1-u_i} \qquad (2)$$

$$W = \sum_{i=1}^{k} w_i \quad (3)$$

The decision rule uses two thresholds. A pair with $W \geq T_u$ is linked automatically, a pair with $W \leq T_l$ is treated as a non-match, and the pairs between the thresholds go to clerical review (Fellegi & Sunter, 1969). Table 2 gives illustrative probabilities for four fields and the resulting weights. A mobile number that agrees is strong evidence, because two unrelated customers rarely share one, and a date of birth that disagrees is strong evidence against a match. Table 3 applies the rule with

T_u = 12 and T_l = 3 to three candidate pairs. Pair 1 agrees on surname, date of birth, and postcode and scores 14.69, so it is linked. Pair 2 agrees only on surname and postcode and scores 0.70, so it is a non-match. Pair 3 agrees on surname and date of birth but disagrees on mobile number and postcode and scores 7.94, which falls in the review band.

Table 2: Probabilities of the illustrative fields and corresponding Fellegi-Sunter weights (in base-2 logarithms).

Field	Agreement among matches m_i	Agreement among non-matches u_i	Agreement weight w_i^+	Disagreement weight w_i^-
Surname	0.95	0.0200	5.57	-4.29
Date of birth	0.98	0.0030	8.35	-5.64
Mobile number	0.90	0.0005	10.81	-3.32
Postcode	0.85	0.0500	4.09	-2.66

Table 3: Decisions for three illustrative pairs at the upper threshold 12 and the lower threshold 3.

Pair	Fields that agree	Fields that disagree	Score W	Decision
Pair 1	Last name, Date of birth, Postcode	Mobile number	14.69	Link
Pair 2	Surname, Postcode	Date of birth, Mobile number	0.70	Non-match
Pair 3	Surname, Date of birth	Mobile number, Postcode	7.94	Clerical review

Weights are base-2 logarithms, and all weights and scores in Tables 2 and 3 are computed from the unrounded m and u probabilities and shown rounded to two decimal places for display; every decision is taken on the unrounded scores, so recomputing a score from the displayed weights can differ from the printed score by a rounding step only. Two features of this rule matter for a multi-unit group. First, the costs of the two error types are unequal. A false merge joins two customers, so that staff serving one may see the other's accounts and a service history is corrupted, which is a privacy incident as well as a data defect. A missed duplicate produces a second record and, at worst, duplicated outreach. Thresholds for automatic linking should therefore be set conservatively and the review band kept wide enough to catch borderline pairs. Second, the probabilities m_i and u_i differ between subsidiaries, because data entry practice and population differ, so they should be estimated again for each business unit from a labelled sample and not copied from the first unit migrated. Christen (2012a) describes the end-to-end matching pipeline of preprocessing, indexing, comparison, classification, and evaluation, and Köpcke and Rahm (2010) compare matching frameworks, which supports choosing and testing the matcher on the group's own data. Elmagarmid *et al.* (2007) survey the field-matching and record-level methods from which the comparison functions are drawn.

Scale makes the comparison cost the critical resource. Comparing all records with each other involves the number of pairs computed by Equation (4), which equals 17,999,997,000,000 or about 1.8×10^{13} in case of six million records. Blocking limits comparison to the records sharing the

same key, while the sorted-neighbourhood method by Hernández and Stolfo (1995) scans a sliding window along records sorted by a key, using several passes with different keys in order to capture pairs separated by some key. Literature surveys on blocking and filtering methods (Papadakis *et* al., 2020) and end-to-end entity resolution for big data (Christophides *et* al., 2021) cover the techniques used. In the case where a learned matcher is used to replace or supplement the weight rule, the maintenance cost described by Sculley *et al.* (2015) for machine learning systems becomes relevant: the matcher relies on data and thresholds that tend to drift, requiring retraining and monitoring of the system.

$$N_{\text{pairs}} = \frac{n(n-1)}{2} \quad (4)$$

3.4 Load design on the target platform

The documentation for Salesforce mentions the capabilities and conditions where bulk loading capacity decreases, and the values differ depending on the API version and thus are provided in this context. For Bulk API 2.0 in version 51.0, ingesting of records allows for no more than 150,000,000 records in the rolling 24-hour period, and the batch limit of 15,000 per rolling 24 hours coincides with that for the older Bulk API; the limit of 150 MB for a single job file and the fact that base64 encoding increases the size of the file are mentioned (Salesforce, n.d.-a). These values are taken from the developer guide on 22 October 2026 and depend on the API version and org edition, so the group should get the live values from the limit resource of its org. Six million records are very far away from the daily record limit. Planning math shows the structure of the job: with a batch size of 10,000 records, six million records will constitute 600 batches, and assuming that the average size of a record is 1 KB, the data will be 6,000 MB, or 40 job files at most, if each raw file is not exceeding 150 MB and a margin is taken into account. The values are not the limitation, but batch structure is.

It is important to differentiate between the two types of values. Daily ingestion allowance, job-file size, behavior during lock waiting, and ownership-skew threshold are the documented constraints of the platform and apply in the form published by Salesforce. Actual throughput of a given org, however, is the environment-specific performance assumption and depends on the edition of the platform, sharing configuration, object complexity, and concurrent activity, and only a rehearsal on the target org will provide the measurement.

There is a difference between the lock behavior of the two APIs. For the earlier Bulk API (version 1), the default is the parallel processing and possible lock contention. According to Salesforce, the API waits a few seconds for the lock and fails the record, retries batches where more than 100 records face lock problem, and tries to run the whole batch up to 10 times (Salesforce, n.d.-a). Using serial mode reduces the possibility of the lock contention but makes the process slower, and the records belonging to the same parent, such as the account team members in one account, should be put in the same batch. For the Bulk API 2.0, the job is divided into batches automatically, parallel execution is used, and failed batches are retried automatically, so the control over the batch structure moves to the order and grouping of the input file and to the job result files that list the failed records (Salesforce, n.d.-a). The appropriate behavior is the one of the chosen API, and the rehearsal should be run on the appropriate API.

Ownership data is the source of the second constraint. Salesforce defines the ownership data skew as one user owning more than 10,000 records of an object and warns that any change in the place of the user in the role hierarchy or his or her group membership will result in very big recalculation

of sharing tables (Salesforce, n.d.-b). The migration assigning every record of the legacy database to one integration user produces exactly such a condition. Possible mitigation is spreading the ownership across many users, avoiding placing the owner of the object in the role hierarchy, and deferring the sharing calculation during the load and performing the recalculation after the end of the load, which means that there is needed a full recalculation of the sharing rules before any changes take place (Salesforce, n.d.-b). According to Salesforce, the appropriate rehearsal should be performed in a full copy sandbox refreshed in the last 30 days. It perfectly fits the recommendations of the Brodie and Stonebraker (1995) staging model: each subsidiary is rehearsed, loaded, and reconciled before the next one starts.

The migration planning in the supplementary collection includes data-integrated ERP migration for the government workforce transformation (Chibwaye & Oteri, 2023), lessons learned from the migration of applications and information systems to cloud infrastructure in a South African retail bank (Ladapo *et* al., 2025), and best practices for cloud migration strategies in the enterprise (Ladapo *et* al., 2022). They are used for the load planning and sequencing and are not CRM studies.

3.5 Reconciliation and statistical validation

Reconciliation involves checking whether what came in equals what went out. Control totals are the traditional tool for this: record counts per item and the sums of a numeric field in the source and the target. These can reveal missing or duplicated records and gross value errors. However, they cannot reveal offsetting errors where two records exchange phone numbers or the account balances of two customers get exchanged, as the control totals are not affected. Key reconciliation plugs part of this gap: with the legacy identifier carried through into the target, each key in the source must correspond to exactly one record in the target and vice versa, and any mismatch on either side is a defect. With entity resolution where duplicates are merged the mapping is not one to one any more; Section 3.18 gives the survivor map which replaces it, so this control is stated in its simple form here and generalised there. Comparison on a sample of fields provides the full control, because this involves checking values rather than identity alone (Pipino *et* al., 2002). Reconciliation is stated here for source records; Section 3.18 distinguishes source-record reconciliation from unique-entity reconciliation when deduplication is performed.

Sampling turns record-level audit into a calculation. For estimating a defect proportion p within a margin of error e at a confidence level with normal deviate z, the standard sample size for a large population is given by Equation (5), and the finite-population correction for a population of N records by Equation (6) (Cochran, 1977). With $z = 1.96$, $p = 0.5$ as the conservative choice, and $N = 6{,}000{,}000$, Table 4 shows that a margin of 5 percentage points needs 385 records, 1 point needs 9,589, and 0.5 points needs 38,172. The correction is almost negligible at this scale, which Figure 1 shows: a wave of 500,000 records needs nearly the same sample as the whole migration. Sample size depends on precision, not on population size, so each stratum that is reported separately, such as each object and each subsidiary, needs its own sample.

$$n_0 = \frac{z^2 p(1-p)}{e^2} \quad (5)$$

$$n = \frac{n_0}{1+\frac{n_0-1}{N}} \quad (6)$$

A clean sample requires careful reading. If an audit of n records finds no defects, the upper 95 per cent bound on the defect rate is approximately $3/n$, the rule of three that Hanley and Lippman-Hand (1983) set out for zero numerators and that is given by Equation (7). A clean sample of 385

records therefore licenses only the statement that the defect rate is below about 0.78 per cent, which may be too loose for identity fields. Audit practice shows that statistical and judgemental sampling coexist: Maingot and Quon (2009) found that internal auditors choose between them and size samples by different rules. For migration, the practical position is to use statistical sampling to establish the defect bound and judgemental sampling to target known risks, such as records from the merged-duplicate set and records near the review band.

Two sampling questions need to be kept separate. Equations (5) and (6) determine the sample size required to estimate an unknown defect rate with the selected margin of error; they address how precisely p is to be measured. Acceptance is a different sampling plan where the sample size must be sufficient for a rate or lower to demonstrate compliance via the upper confidence bound, which is a one-sided statement on exceeding a threshold, and not a two-sided statement on precision. Two different designs result in different sample sizes, because the former depends on estimation precision and the latter on the difference between the expected rate and the threshold and the choice of one-sided or two-sided bounds.

$$p_{\mathrm{up}} \approx \frac{3}{n} \quad (7)$$

Table 4: Required sample size for auditing at 95 per cent confidence with = 0.5, and the upper confidence bound on the defect rate assuming no defects in the sample.

Margin of error	Large-population size n_0	n for N = 6,000,000	n for N = 500,000	Upper bound if no defects, 6,000,000
5.0 points	385	385	384	0.779 per cent
2.0 points	2,401	2,401	2,390	0.125 per cent
1.0 points	9,604	9,589	9,424	0.031 per cent
0.5 points	38,416	38,172	35,676	0.008 per cent

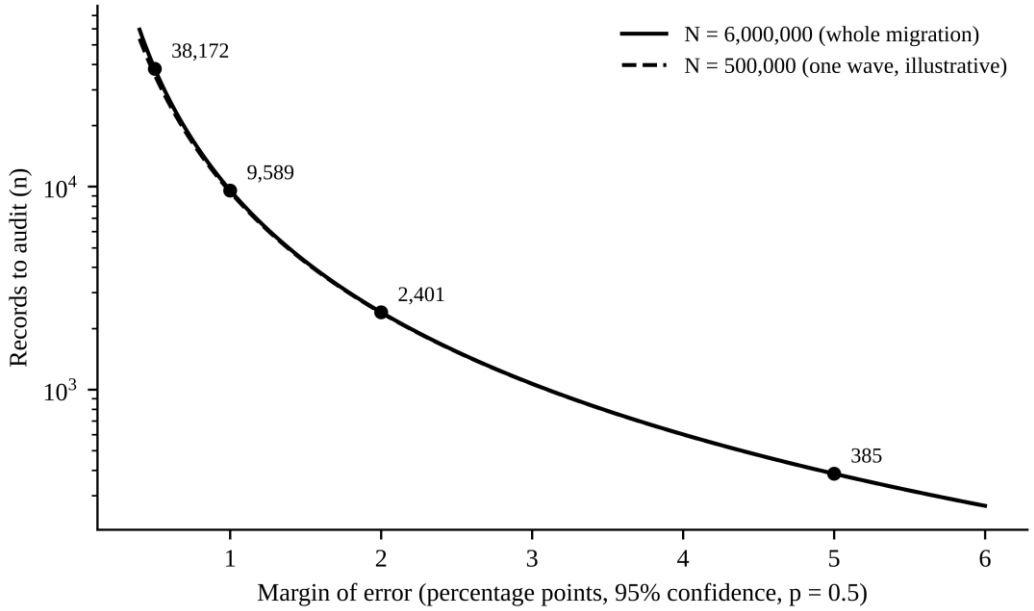

Figure 1: Required audit sample size for the entire migration and for an illustrative migration wave of 500,000 records against margin of error. Finite-population correction in Equation (6) is applied to the sample sizes.

3.6 Security, privacy, and compliance during migration

A migration will copy personal data to new databases, and each database inherits responsibilities. Regulation (EU) 2016/679 requires personal data to be accurate, necessary, limited in storage period and protected from unauthorised processing (European Parliament & Council of the European Union, 2016). All these require migration consequences. Accuracy requires the implementation of the cleansing gate, as inaccurate records should not be copied. Minimisation requires that inactive records and fields unused by consumers are not migrated. Storage limitation means staging extracts and quarantine tables must be deleted on a regular basis after signing off of a wave. Security requires that extracts in transit and in storage be secure and access to staging areas restricted to the migration team.

For the Nigerian subsidiaries of a group, the Nigeria Data Protection Act, 2023, assented to 12 June 2023, imposes obligations on controllers and processors of personal data in Nigeria (Federal Republic of Nigeria, 2023), and a consolidation that transfers data across borders or onto a shared platform requires a documented lawful basis and transfer arrangement according to the law for each source. Non-production environments must be attended to in the same way. Because a full-copy sandbox contains a copy of production data, the rehearsal that Salesforce recommends before postponing sharing calculations (Salesforce, n.d.-b) produces another environment with production data confidentiality requirements, which calls for masking or restricted access (European Parliament & Council of the European Union, 2016).

Privacy and compliance issues during migration are covered in a conceptual framework for harmonising data protection compliance among GDPR, NDPR, and United States privacy regulations (Ominyi & Anichukwueze, 2022), data privacy governance models for cross-border digital platforms (Annan, 2022), a privacy-preserving data architecture for regulated analytics under GDPR and HIPAA (Atakpa & Abolaji, 2022), and evaluations of data protection impact

assessment models in multi-cloud financial infrastructure (Mbonu *et* al., 2022) and comparative studies of data protection regulations and cloud implementation in secure cloud computing (Mbonu *et* al., 2019). Multi-jurisdictional banking post-merger compliance integration (Ominyi, 2020) is another description of the same exposure when systems are consolidated.

3.7 An integrated five-gate framework

The analysis produces five gates, depicted in Figure 2. In each case, the wave, which is the data records of one subsidiary or one group of objects, flows through each of the five gates in sequence, and failing any exit criterion means the wave reverts to an earlier gate. Gate 1 performs a profile on each source and registers defect classes with the data owner (Kim *et* al., 2003). Gate 2 cleanses the data by rule, quarantines data that cannot be fixed by rule, and compares the quality score of Equation (1) against a floor value for each mandatory field (Pipino *et* al., 2002). Gate 3 resolves entities using blocking, the match weight of Equations (2) and (3), two thresholds, and a clerical review band, estimating probability values for each unit (Fellegi & Sunter, 1969). Gate 4 loads data in batches with batch sizes, owner distribution, and deferred sharing set by the platform limits and rehearsed in a full-copy sandbox (Salesforce, n.d.-a; Salesforce, n.d.-b). Gate 5 reconciles data by counts, key cross reference, and record-level auditing of size determined by Equations (5) to (7) (Cochran, 1977).

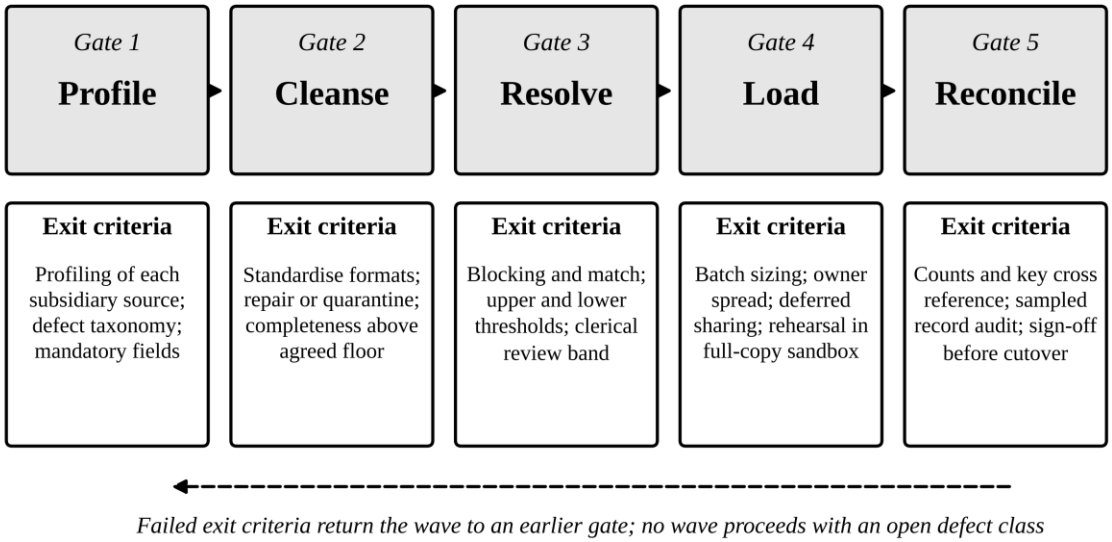

Figure 2: The five validation gates in a staged CRM migration process, with exit criteria and a return path for unsuccessful waves. The return path differentiates the failed wave that re-enters an earlier gate for automated reprocessing from the quarantined and merged records that need manual approval by the data owner.

Table 5 shows, for each gate, the control measures, the measurement, and the exit criterion proposed. These criteria are suggestions for a group to calibrate, and the numerical floors shown are examples of the form the criteria take and not benchmarks drawn from the literature. What the conceptual framework establishes is the sequence of the controls, the necessity that there be a criterion for each control, and the rule that the wave does not move forward with any open defect class.

Table 5: Gates, controls, metrics, and exit criteria proposed for staged CRM migration

Gate	Controls	Metric	Proposed exit criterion
1. Profile	Column profiling per source; defect taxonomy (Kim *et al.*, 2003); map of mandatory fields to target	Share of source columns profiled; defect classes registered	Every column profiled and mapped; defect register accepted by the data owner
2. Cleanse	Rule based standardization; Quarantine of files requiring judgment (Rahm & Do, 2000)	Quality score Q per mandatory field (Pipino *et al.*, 2002)	Q at or above the agreed floor for each field (for example 0.99 for identifiers); quarantine queue owned and dated
3. Resolve	Blocking; Fellegi-Sunter weights; Two thresholds; Clerical check range (Fellegi & Sunter, 1969; Papadakis *et* al., 2020)	Share of pairs in review band; false-merge rate in audit sample	Parameters estimated for the unit; false-merge rate below the agreed bound in a sample sized by Equation (5)
4. Load	Batch sizing; owner spread; deferred sharing; sandbox rehearsal (Salesforce, n.d.-a; Salesforce, n.d.-b)	Batch failure and retry rates; lock errors; owners above 10,000 records	Rehearsal completed in a full-copy sandbox with the measured batch failure rate at or below the agreed ceiling (for example 0.5 per cent) and no unapproved ownership skew after remediation
5. Reconcile	Counts and control totals; key cross reference; sampled field comparison (Cochran, 1977; Hanley & Lippman-Hand, 1983)	Unmatched keys; control-total differences; sample defect rate and upper bound	Zero unexplained differences; every key mapped once; defect bound within tolerance

Interoperability and quality frameworks for health systems (Ilodigwe & Adesemoye, 2021), cloud interoperability model for secure exchange of health information between organizations (Osinem *et* al., 2024), comparative frameworks for verification, risk assessment, and confidentiality in health information governance (Peteranaba *et* al., 2022), and research on standardization of unstandardized compliance controls in high-volume distribution networks (Uchechi *et* al., 2023) support the use of a gated structure proposed above. However, they are mentioned as an illustration of the design principle of explicit controls at handover points and not as validation of the framework.

Hence, the claim of novelty of the framework is not any particular control, since each of the gates is based on existing practice, but the bundling of controls in a gated sequence with exit criteria, parameters to be estimated for each business unit, and the reconciliable identity. Such a sequence

is more preferable than other possible control structures for two reasons. First, controls are ordered in the way that each depends on the completion of the previous one: profiling determines defect classes that are corrected in cleansing, and the identity hygiene is provided by the Gates 1 and 2 and is a prerequisite for entity resolution and reconciliation, therefore a structure that works with data prior to identity resolution will not reuse the same evidence. Second, the gated structure requires an explicit decision to be made at each handover point, and thus will preserve the information about failure attributions that would be lost in programs relying on end-of-project testing. Either a flat structure or one ordering cleansing after the resolution of identities will lose one of the two.

Table 6: Positioning of the five-gate framework relative to the existing approaches it relies on.

Established approach (representative sources)	**Capability provided**	**Limitation for CRM migration control**	**What the five-gate framework adds**
Data quality assessment methodologies (Pipino *et al.*, 2002; Batini *et al.*, 2009)	Metric forms for quality dimensions; assessment and improvement cycles	Metrics are not tied to load or cutover decisions and carry no acceptance rule	Operationalises the metrics as gate exit criteria with owners and floors
Record linkage and entity resolution theory (Fellegi & Sunter, 1969; Christen, 2012b; Christophides *et al.*, 2021)	Decision rule for pairs; blocking; evaluation measures	Stops at classifying pairs; no wave-level acceptance decision	Ties thresholds, the clerical band, and a linkage audit to a wave go or no-go
Incremental legacy migration strategy (Brodie & Stonebraker, 1995)	Staged retirement of legacy systems	Predates probabilistic entity resolution and modern platform load limits	Embeds the gates and per-unit parameters in the staged waves
Data governance frameworks (Khatri & Brown, 2010; Abraham *et al.*, 2019)	Decision domains, roles, and accountability	Define ownership but not operational acceptance criteria	Assigns each gate an accountable owner and a measurable criterion
ETL and warehouse practice (Kimball & Caserta, 2004; Vassiliadis, 2009)	Pipeline design, idempotent loads, staging	Reconciliation limited to control totals	Adds sampled, interval-based reconciliation and a survivor-map identity

Established approach (representative sources)	Capability provided	Limitation for CRM migration control	What the five-gate framework adds
Platform vendor guidance (Salesforce, n.d.-a; n.d.-b)	Documented limits, skew mitigation, rehearsal advice	Product-scoped and version-dependent; no cross-platform synthesis	Generalises platform facts into gates with versioned, sourced limits

3.8 Critical appraisal of the evidence base

Foundations of the field are well-established and sound. Multidimensional approach to data quality (Wang & Strong, 1996), probabilistic record linkage (Fellegi & Sunter, 1969) and the theory of sampling (Cochran, 1977) have stood the test of time and have been improved but not superseded (Batini *et* al., 2009; Christophides *et* al., 2021). The weakness is in the connections. Information on defects in enterprise data systems is available only through a few case studies (Xu *et* al., 2002; Zahari *et* al., 2015), while the literature on CRM implementation is focused on strategic, organizational and financial aspects of the process (Chen & Popovich, 2003; Krasnikov *et* al., 2009). There are no empirical studies validating gate-based migration control frameworks against the defect outcomes.

Evidence from the Platform Evidence of this sort is authoritative for the platform, but evolves with releases, and the numbers cited here are as seen on 22 October 2026 (Salesforce, n.d.-a; Salesforce, n.d.-b). There are three implications. The gate should be interpreted as a synthesis of sound parts, not a verified process. The numeric criteria are to be used locally. And the most valuable evidence that the field could bring in would be comparative: migrations that documented gate results and subsequent production defect rates would show which gates predict problems.

3.9 Measuring data quality and automating verification

Measurement reduces quality dimensions from Section 3.1 to numeric values that are testable with a gate. Literature on data and information quality has been structured around its topics and methods (Madnick *et* al., 2009), and research on the Big Data era has pointed out how volume, variety, and velocity change the picture (Cai & Zhu, 2015). A tutorial on data quality from theory to practice discusses consistency, deduplication, currency, and accuracy as separate technical problems with separate algorithms (Fan, 2015). From the research on CRM migration the most important finding is that the dimensions are not mutually substitutive: a table can be complete and incorrect, or correct and duplicated, so a gate that tests one dimension says nothing about the others. ISO/IEC 25024 specifies measurements for the data quality characteristics listed in ISO/IEC 25012 (ISO/IEC, 2015), giving the team a terminology for writing acceptance criteria recognizable to the auditor.

Profiling is the first step in measurement. A review of profiling techniques for relational data covers tasks from single-column statistics to dependency discovery and emphasizes that profiling helps both cleansing and schema mapping (Abedjan *et* al., 2015). What counts for a CRM load is practical: the null rate, distinct count, and pattern distribution of every field that will be mapped; the cardinality of identifier fields; and the distribution of values in picklist fields that must map to the target's controlled vocabulary. Verification by automated means comes next by declaring

constraints and verifying them at scale. An approach to automated verification of large-scale data quality constraints uses declarative specification of constraints and performs large-scale evaluation with incremental computation (Schelter *et* al., 2018), which is the right model for performing the same checks after each migration wave. A classification of ERP data quality problems separates master and transaction data quality problems (Haug *et* al., 2009), while a research on data migration highlights that quality assessment and cleansing are prerequisites for migration (Azeroual & Jha, 2021). A case study of a CRM implementation showed the data quality problems that had emerged only in the system's operation (Reid & Catterall, 2005), making a case for testing not just the format of fields but their contents.

Table 7: Data Quality Dimensions, Example Measures for a CRM Migration, and Gate at Which They Are Tested

Dimension	Example measure	Formula or test	Gate
Completeness	Share of mandatory fields populated	Q = 1 - N_missing / N_required	Profiling; reconciliation
Uniqueness	Share of customers represented once	Distinct entities / records after resolution	Entity resolution
Validity	Share of values that meet format and domain rules	Rule pass count / rule test count	Cleansing
Consistency	Share of related fields that agree across objects	Cross-object rule pass rate	Cleansing; reconciliation
Accuracy	Share of sampled records that match the source of truth	Sample proportion with Wilson interval	Reconciliation
Referential integrity	Share of child records with a valid parent	R = 1 - N_orphan / N_child, with N_orphan the child records whose parent is missing or invalid	Load

In Table 7 six dimensions are presented with example measures and the gate at which each dimension is to be tested. The formulas are generic and the group is supposed to write down its own definitions prior to the first migration wave, to make results comparable among business units. The uniqueness ratio of Table 7 is a necessary but not sufficient indicator of entity-resolution accuracy: a false merge and a missed duplicate can balance each other and leave the ratio intact, so it is to be interpreted together with the audited false-merge and missed-duplicate rates of Gate 3 and with the evaluation measures of Section 4.8.

3.10 Entity resolution at scale: blocking, evaluation, and learning

Entity resolution has three technical issues that affect each other. They are minimizing the number of candidate pairs, choosing matching pairs among candidates, and evaluating the results. The number of candidate pairs is proportional to the square of record counts, and thus, all 1.8×10^{13} pairs of six million records would have to be compared. Indexing and blocking methods minimize

the comparisons of only those records that share the blocking key, and a review of indexing approaches for scalable record linkage and deduplication evaluates them in terms of reduction and completeness (Christen, 2012b). An experimental evaluation of approximate blocking techniques shows that the results vary greatly depending on the data and parameters setting (Papadakis *et* al., 2016). A tutorial on theory, practice, and challenges of entity resolution provides a pipeline overview (Getoor & Machanavajjhala, 2012). A generic framework considers match and merge operations as black boxes and provides algorithms that achieve the same result through fewer comparisons (Benjelloun *et* al., 2009). An introductory text on duplicate detection covers similarity measures and evaluation (Naumann & Herschel, 2010).

There are three performance metrics of a blocking scheme. Let T be the number of true matching pairs, C be the set of candidate pairs found after blocking, and C_T be the set of true matches in the set of candidates. Reduction ratio, pairs completeness, and pairs quality are defined by Equations 8 to 10.

$$RR = 1 - \frac{|C|}{n(n-1)/2} \qquad (8)$$

$$PC = \frac{|C_T|}{T} \qquad (9)$$

$$PQ = \frac{|C_T|}{|C|} \qquad (10)$$

Let us consider a blocking scheme for the illustrative example of six million records. Assume that the blocking key generates 1.5 million blocks of 4 records on average. The number of candidate pairs generated in each block equals 6, and thus, the candidate set contains 9.0 million pairs of the total 1.8×10^{13} possible pairs, RR equals 0.9999995. Assume also that the blocking key leaves 4,700 of 5,000 audited true matches in the same block, and thus, PC equals 0.94, that is, 6 per cent of true duplicates will not be compared at all. Since the pairs completeness constrains the recall, the usual remedy to overcome this limitation is applying another blocking scheme based on a different key, like normalized mobile number, and the gate has to report both the value of PC and the match thresholds.

The following three keys illustrate the trade-off between them. The key composed of a phonetic surname code and a postcode prefix can generate 1.5 million blocks of 4 records resulting in 9,000,000 candidate pairs. The key composed of a normalized mobile number can generate 1.8 million blocks of 2 records resulting in 1,800,000 candidate pairs but only for customers having the phone number. Thus, the latter key provides a poor completeness for the remaining customers. The key composed of the birth year and the first initial can generate 150,000 blocks of 40 records resulting in 117,000,000 pairs, which still reduces the number of possible pairs by 99.9993 per cent from 17,999,997,000,000 but is much more than the former key and produces many non-matching pairs. These block sizes are assumed only for the illustration purpose. The lesson is that the keys should be combined into a union since each one finds matches missing for the others. The blocking scheme must limit the block size in order to avoid the problem when the same surname occurs frequently in the same postcode (Christen, 2012b).

Evaluation of the quality of linkage requires special care since matched and unmatched pairs are highly imbalanced. A guide for researchers analyzing the linked data explains that the false matches and missed matches distort the analysis and provides a solution to estimate the linkage error and report it (Harron *et* al., 2017). Applied probabilistic linkage in epidemiology reviews the

choice of match-weight thresholds and the review of the grey zone (Sayers *et* al., 2016). A fast estimator generalizes the Fellegi-Sunter model for administrative files with millions of records (Enamorado *et* al., 2019). A book on data quality and record linkage techniques covers editing, imputation, and matching methods in detail (Herzog *et* al., 2007). Learned matchers become competitive: pre-trained language models used for entity matching improved the results on benchmark datasets (Li *et* al., 2021). For the CRM migration the practical implication is that learned matchers can improve the recall on messy names and addresses, but they require labeled pairs from each business unit and the calibrated threshold, so the clerical review band of Section 3.3 becomes necessary.

3.11 Cleansing, repair, and ETL design

Automated repairs should be as explicit as matches. Monograph on trends in cleaning relational data discusses constraint-based repair and deduplication and indicates that automated repair decisions might be erroneous if the set of constraints is incomplete (Ilyas & Chu, 2015). System for holistic data repair uses probabilistic inference for combining signals including constraints, external data, and statistics (Rekatsinas *et* al., 2017) – and it is an example of the potential but also need for human validation of such an approach. The adopted rule of automating repairs according to deterministic rules and flagging other values is in line with this precaution.

The design of extract, transform, load pipeline determines whether the pipeline can be run safely again. Survey of ETL technology categorizes approaches to ETL by modeling, design, and optimization of the process (Vassiliadis, 2009), and practice-oriented monograph provides guidelines for extraction, cleaning, conformation, and delivery of data (Kimball & Caserta, 2004). Three properties of design apply to the pipeline of CRM migration. The pipeline should be idempotent to be able to rerun a failed batch without creating duplicates; the pipeline should move the legacy identifier to an external identifier field in each target record for reconciliation and rollback purposes; and data should be staged to allow testing transformations before writing to the target. The practice literature reaches similar conclusions regarding reconciliation controls and redesigning of financial workflows (Dogbatsey *et* al., 2019) and multi-enterprise ETL architecture with Lean Six Sigma guardrails (Ogundairo, 2025). Technical due diligence and supply chain integration in the energy sector touches on system integration following acquisitions (Nwangwu *et* al., 2023).

3.12 Interval estimates for audit results

An audit sample gives a proportion, and a proportion should have an interval. Comparison of interval estimators for a binomial proportion indicated that the textbook Wald interval has poor coverage, especially in case of a small sample size or proportion close to zero or one, and proposed alternatives such as Wilson interval (Brown *et* al., 2001). The Wilson interval was derived from the score test (Wilson, 1927), and it has the additional advantage of being informative even when no defects are discovered. For k defects in n audited records, with proportion p = k/n, and with normal quantile z, the Wilson interval is calculated as shown in Equation 11.

$$\frac{p+\frac{z^2}{2n}\pm z\sqrt{\frac{p(1-p)}{n}+\frac{z^2}{4n^2}}}{1+\frac{z^2}{n}} \qquad (11)$$

Table 8 demonstrates the use of the Wilson interval for five audit cases. Count is illustrative, but the key point is to demonstrate that the decision is driven by the upper bound and not by the point estimate. In case of acceptance criterion where the upper bound of 95 percent confidence interval

on the record defect rate must be less than 1 percent, Subsidiary C is accepted with zero defects in 1,100 records, while Subsidiary D, despite having a point estimate of 0.80 percent, which looks acceptable, is not accepted since its upper bound is above 1 percent.

Table 8: Illustrative audit results and Wilson 95 percent intervals for five subsidiaries against a 1 percent upper-bound criterion.

Subsidiary	Records audited	Defects	Defect rate (%)	Lower bound (%)	Upper bound (%)	Decision
Subsidiary A	1,500	3	0.20	0.07	0.59	Pass
Subsidiary B	1,500	7	0.47	0.23	0.96	Pass
Subsidiary C	1,100	0	0.00	0.00	0.35	Pass
Subsidiary D	1,500	12	0.80	0.46	1.39	Review
Subsidiary E	900	2	0.22	0.06	0.81	Pass

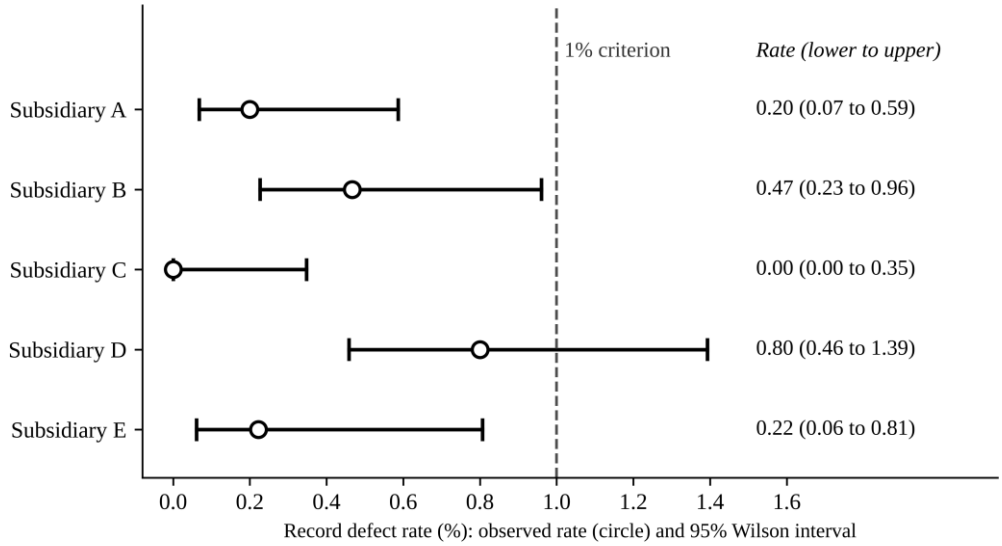

Figure 3: Observed defect rates and 95 per cent Wilson intervals for the five illustrative audit results, with the 1 per cent acceptance criterion. Intervals are two-sided 95 per cent Wilson intervals, and bounds are shown to two decimal places.

Figure 3 illustrates the results graphically. The point estimate for Subsidiary D is to the left of the criterion, but the interval crosses it, which is the reason the decision hinges on the upper bound.

Acceptance criteria have been defined as follows. The criterion in Gate 5 of Table 5 uses the upper endpoint of a two-sided 95 per cent Wilson interval, which corresponds to the upper bound of a one-sided interval at 97.5 per cent confidence level. A one-sided 95 per cent bound using normal deviate 1.645 would produce narrower intervals and smaller sample sizes, but the criterion must be fixed before the audit, not selected in view of the observed result. Sample size computations in Equations (5) and (6) provide precise estimates, but in themselves cannot prove compliance with a maximum acceptable rate, which is the task of the upper bound.

There is a parallel in audit sampling for accounting research. Audit sampling research literature synthesis touches upon the choice of statistical versus non-statistical procedures and the research on auditor judgement (Elder *et* al., 2013), and the finite-population correction applied in Section 3.5 is identical to the adjustment made in attribute sampling. Clear audit sample proves a bounded defect rate, not a zero defect rate, and that is the meaning of the acceptance statement.

3.13 Privacy and security controls for migration data

Migration makes duplicates. Extracts, staging tables, error files, sandbox refreshes all may include personal information, and each carries the responsibility of Section 3.6. The common belief that anonymisation happens once the names are stripped from personal data is incorrect because k-anonymity is defined as a data release where each record looks indistinguishable from at least $k - 1$ others based on the quasi-identifiers (Sweeney, 2002), and l-diversity demonstrates that k-anonymity in itself is not sufficient protection of sensitive attribute values within homogeneous groups (Machanavajjhala *et* al., 2007). For test and training data it implies that the masking procedure should remove links to actual customers and should be checked for re-identification risk. The NIST Privacy Framework provides the framework for managing risks of privacy impacts throughout the lifecycle of data (National Institute of Standards and Technology, 2020), and the analysis of the European Union General Data Protection Regulation spells out its requirements for organisations processing personal information (Tikkinen-Piri *et* al., 2018). A study on the effect of GDPR on database systems concluded that features such as timely deletion and metadata indexing incur performance overhead (Shastri *et* al., 2020), and that is relevant to the sizing of migration window and the erasure after load.

The practices discussed in the collection deal with the governance of health data for privacy preservation using blockchain and cryptography-based approaches (Afrihyia *et* al., 2024), the governance of data in digital health learning systems (Orise & Niniola, 2025), the integrity and governance of data in regulated life sciences operations (Aneke *et* al., 2023), and privacy-focused security engineering in the cloud systems (Okoruwa *et* al., 2020). Legal and ethical risk modelling for the governance of enterprise data protection (Mbonu *et* al., 2018), data lakehouse governance for data loss prevention (Aliliele *et* al., 2025), and modernisation of the enterprise firewall for business continuity during secure migration (Ogunwola & Alozie, 2026) tackle control issues which are close to the subject. Interoperability models for distributed ledger payment systems (Adesuyi *et* al., 2023) are included due to the stated topic relevance, and they indicate that the migration of integration layers is a separate workstream from data migration.

3.14 Governance, ownership, and the risk register

Governance decides which entity is allowed to assume a risk. A systematic literature review of data governance provides a conceptual framework which consists of scope, mechanisms and outcomes (Abraham *et* al., 2019), and the earlier designed conceptual framework for data governance includes decision domains for the principles, quality, metadata, access and lifecycle of data (Khatri & Brown, 2010). Systematic literature review of CRM research highlights implementation problems including data and organisation-related issues (Boulding *et* al., 2005). In the case of migration, those ideas translate into a risk register. The risk register is presented in Table 9 with the gate which addresses each risk, control, and an accountable owner. The idea behind the risk register is to have it considered at every gate and not only at the start since usually the first wave identifies underestimation of some risks. If some risk reoccurs during different

waves, it needs to be moved from the programme risk register to the ongoing data governance agenda with an owner and a measurement which persists beyond migration.

Table 9: Migration risk register with gate, control, and accountable owner.

Risk	Gate	Control	Owner
Duplicate customers across subsidiaries	Entity resolution	Blocking with two keys; threshold per business unit; clerical review band	Group data owner
False merge of distinct customers	Entity resolution	Upper threshold set from cost of false merge; manual approval of merges above a risk tier	Business unit data owner
Invalid or non-standard values	Cleansing	Deterministic repair rules; quarantine of unrepairable rows	Migration lead
Record lock contention and failed batches	Load	Batch sizing; serial processing for parent objects; retry with idempotent keys	Platform administrator
Ownership skew after load	Load	Distribute record ownership across users or queues	Platform administrator
Masking off-set errors through control totals	Reconciliation	Record-level audit sample with interval estimates	Quality assurance lead
Personal information in staging and errors files	All	Encryption, access limits, retention schedule, masked test data	Data protection officer
Failed cutover	Load; reconciliation	Rollback plan using legacy identifiers; dress rehearsal	Programme sponsor

3.15 A worked wave plan

Each wave needs to be planned as it ties the gates to the time frame. It is assumed that six million records are stored in five subsidiaries in proportions described in Table 10, and that the loading is done at the rate of 40,000 records per minute. Load time for a wave equals to the number of records divided by the rate, while the elapsed window is planned twice as long as the load time to accommodate two-pass loading of parent and child entities, retries and post-loading reconciliation. Those are planning figures for illustration only. Waves go by the order of risks – from the smallest and cleanest to the largest and messiest subsidiary in order to test the process on four waves before the toughest one (Sayers *et* al., 2016; Harron *et* al., 2017).

Table 10: Illustrative wave plan at 40,000 records per minute with a planning window of twice the load time.

Wave	Records	Load time (minutes)	Planned window (minutes)	Planned window (hours)
Subsidiary C	900,000	22.5	45.0	0.8
Subsidiary E	800,000	20.0	40.0	0.7
Subsidiary B	1,400,000	35.0	70.0	1.2
Subsidiary D	1,000,000	25.0	50.0	0.8
Subsidiary A	1,900,000	47.5	95.0	1.6
Total	6,000,000	150.0	300.0	5.0

From Table 10, we see that the longest wave, Subsidiary A, will need an estimated planned window of about 1.6 hours, and the whole program will need about 5 hours of estimated load windows. This is relatively short when compared to the preparatory work needed. What poses more risk is not the load itself but the number of gates. Each wave will have to pass through five gates, and hence the program will pass through 25 gates in all. Therefore, the criteria listed in Table 5 must be determined before the first wave to avoid adjustments based on results after they occur.

An estimated average of 40,000 records per minute is just a planning assumption and not the platform limit. Such an estimated average is consistent with bulk loads of simple parent objects in an org with deferred sharing and low concurrent activity. However, the actual achieved average will depend on edition, org load, and object complexity, and this can only be verified through dress rehearsal. This plan is independent of the assumed rate in terms of shape. At half the rate, the total planned window is doubled, to about 10 hours, still fitting into overnight maintenance window with some room to spare, and at double the rate it is cut in half, to about 3.3 hours. Hence, the wave order, gate sequence and reconciliation design remain the same within this range. The assumption should be replaced with the actual rate measured in the dress rehearsal rather than refined in advance.

3.16 Monitoring data quality after go-live

Quality does not end at cutover. Declarative checks such as those used in large-scale verifications (Schelter *et* al., 2018) may be scheduled to run against live CRM, and measures of Table 7 may be monitored as indicators. Hidden dependencies in data pipelines incur technical debt over time (Sculley *et* al., 2015). This is why the checks used by a migration team must be left behind, not thrown away during handoff. Data decay is yet another challenge. Customer contact data will change with varying market pace and, therefore, the currency aspect of quality (Fan, 2015) has to be watched. Table 11 gives the list of indicators to monitor CRM data quality in the first year after cutover with a recommended frequency of review.

Table 11: Indicators for monitoring CRM data quality after cutover.

Indicator	Definition	Review frequency	Owner
Duplicate rate	Share of customer records that resolve to an existing customer on a scheduled match run	Monthly	Group data owner
Mandatory field completeness	Share of mandatory fields populated on records created in the period	Weekly	Business unit data owner
Validity rule failures	Count of records failing validation rules after creation	Weekly	Platform administrator
Orphan rate	Share of child records without a valid parent	Monthly	Platform administrator
Sampled accuracy	Defect rate in a random sample with a Wilson interval	Quarterly	Quality assurance lead
Contact data currency	Share of contact records verified in the last twelve months	Quarterly	Business unit data owner

3.17 Stratified audit design across subsidiaries

A single random sample of the pooled records will be dominated by the largest subsidiary and will miss out on the smallest one. Stratified sampling considers each subsidiary as a stratum and allocates sample sizes such that the estimate for each is usable and the estimate for the pool is efficient (Cochran, 1977).

Proportional allocation gives each stratum a share equal to its share of records, whereas Neyman allocation also weights each stratum by its variability, so that strata with a higher expected defect rate receive more of the sample. For a stratum h with N_h records and prior defect proportion p_h, the standard deviation of a record-level indicator is $S_h = \sqrt{p_h(1 - p_h)}$, and the allocation of a total sample n is given by Equation 12.

$$n_h = n \frac{N_h S_h}{\sum_k N_k S_k} \quad (12)$$

The application of both allocation schemes to the total audit budget of 6,000 records, assuming some illustrative values of prior defect probabilities, is shown in Table 12 below. The two allocations differ most for those subsidiaries whose prior defect probabilities are extreme. In order to make the interval narrow enough for every subsidiary to use, the floor of 600 records per stratum is applied. This increases the total allocation somewhat above the budget, but the floor level is a design choice to be set according to the required interval width based on the Wilson interval discussed in Section 3.12 (Brown *et* al., 2001).

Table 12: Illustrative allocation of a 6,000-record audit across subsidiaries under proportional and Neyman rules.

Subsidiary	Records	Prior defect rate (%)	Proportional allocation	Neyman allocation	Neyman with 600 floor
Subsidiary A	1,900,000	0.3	1900	1468	1468
Subsidiary B	1,400,000	0.8	1400	1762	1762
Subsidiary C	900,000	0.2	900	568	600
Subsidiary D	1,000,000	1.0	1000	1405	1405
Subsidiary E	800,000	0.5	800	797	797

Both allocations in Table 12 are calculated via Equation 12 for the stated population sizes and priors and rounded to the nearest record, with stratum weights 103,911; 124,718; 40,209; 99,499; and 56,427 for Subsidiaries A to E respectively and their sum equaling the 6,000-record audit budget before the floor application. The stratification can also be applied across objects so that accounts, contacts, opportunities, and cases receive a share and across the field types of highest business impact, such as identifiers, contact channels, and consent flags. All such design should be pre-defined before the audit starts, with reporting of its results by stratum and calculation of the pooled estimate using stratum weights and not by simple averaging.

3.18 Reconciling identifiers when duplicates are merged

The assumption of key-based reconciliation is that each source key points to a single target record. In contrast, the entity resolution explicitly does not assume the one-to-one relationship between source keys and target records and when two legacy records are merged into a single record describing the same customer, there will be extra target records relative to source records. The number of the extra target records is the number of merges, and the reconciliation that insists on the one-to-one mapping will classify the merged records as lost records. Suppressing the alarm can lead to suppression of the actual losses, too. To address the problem, one should reconcile against the survivor map, recording for each source key the target record that survives and the reason. This way, the mapping becomes many-to-one on purpose, and all source keys are accounted for (Fellegi & Sunter, 1969; Pipino *et* al., 2002).

In terms of the accounting identity of the wave, the N_source = N_target + N_merged + N_quarantined + N_rejected, where N_merged is the number of source records merged into surviving records, N_quarantined is the number of records quarantined for future repair and N_rejected is the number of source records that could not be loaded and were placed into the job result files. Thus, for a wave with 1,000,000 source records producing 962,000 target records, 31,000 merged records, 4,000 quarantined and 3,000 rejected records, the identity holds: 962,000 + 31,000 + 4,000 + 3,000 = 1,000,000. These numbers are illustrative, but any deviation from zero indicates an unexplained difference that prevents the closure of the exit criterion of Gate 5 in Table 5. The survivor map also allows carrying the legacy identifiers of absorbed records into a surviving record, e.g. as a multi-value field or a related cross-reference object. It allows resolving legacy keys used on documents and quotes provided by customers and splitting a wrongly merged pair in the audit without searching in the source. Checking of the map also allows direct testing of the false-merge rate since every merge decision is a record that can be sampled.

4. DISCUSSION

4.1 Implications for migration leads

Three best practices apply to managers of the migration. First, determine acceptance criteria with the users of the data and quantify these criteria in terms of metrics such as the quality score of Equation (1) – as any criterion which cannot be quantified cannot gate the wave (Wang & Strong, 1996). Second, think of entity resolution as a process with its own parameters, labelled sample, and review queue for each business unit, and make the automatic linking conservative as the error of merging two different entities is costlier than the error of failing to match a duplicate (Fellegi & Sunter, 1969). Third, plan load as an engineering task, accounting for known limits of the platform, incorporating the owner spread, deferred sharing, and rehearsals as the platform documentation recommends (Salesforce, n.d.-b).

Reconciliation is to be designed before the load and not after it. The legacy identifier has to carry across into the target system, the audit sample size has to be determined based on the required precision, and the interpretation of a clean sample has to be decided in advance, since a sample of a few hundred records allows for only a broad estimate of the defect rate (Hanley & Lippman-Hand, 1983); the complete interval estimates are presented in Sections 3.5 and 3.12, and the choice of the required precision and of what a clean sample can guarantee lies with the data owner and not with the migration team.

4.2 Implications for data governance and ownership

The dimension of people and process Chen and Popovich (2003) noted applies to migration just as much as to the CRM that comes after it. Each subsidiary has to have a named data owner who takes responsibility for the defect register, determines the status of quarantined records, and authorizes the reconciliation, and the whole group has to have a steward who takes ownership of the matching rules after the migration is done. Without these roles, the thresholds that have been determined during the project become fuzzy, and the single customer view gets degraded as new records come in through five channels. Regulated activities reinforce the point: accuracy and storage limitation requirements of Regulation (EU) 2016/679 (European Parliament & Council of the European Union, 2016) and obligations of controllers of personal data under the Nigeria Data Protection Act, 2023 (Federal Republic of Nigeria, 2023) apply to named individuals and do not end when the migration is finished.

4.3 Limitations

This is a critical review of the selected sources, and its conclusions depend on that selection. The worked examples use illustrative probabilities and planning assumptions, including the average record size; the examples demonstrate calculations and not the actual performance. The limits of the platform that were used may change, and the load discussion does not include every platform feature that may be used in a large migration. The framework has not been applied to migration in practice, so the cost in time and effort is not known. The sample size formulae assume simple random sampling, and stratified or clustered designs would change the calculations (Cochran, 1977).

4.4 A cutover and rollback plan

Cutover is where gates meet the calendar. Dress rehearsal in a sandbox which resembles production, with masking of the data and the same batch sizes gives the schedule and error profile that the live load will exhibit. The plan should include the load window, freeze on changes in the

legacy system, criteria for go/no-go decision, and the rollback procedure. As the legacy identifier carries over to every target record, the rollback can be executed by deleting the records which have been added in the wave only and not the records entered by the users. The same identifiers enable reconciliation of the final state and the audit sample. Migration studies in regulated environments, such as migration of applications and information systems to cloud infrastructure in a South African retail bank (Ladapo *et* al., 2025) and best practices for cloud migration in the enterprise (Ladapo *et* al., 2022), reinforce the point about the dress rehearsal and phased migration.

A rollback that removes only created records is enough for the first loading of an empty object, and is not enough if the wave updates records that already existed, touches records with dependants, or fails mid-way through. There are four different scenarios that require different treatment. In the first scenario, the load modifies existing records: before loading, a before image of all fields the wave is going to update is taken and stored for future use, and for rollback the same is applied as an update, which returns the values, but cannot undo the history entries and system modification stamps that are put there by the load. In the second scenario, there are dependants: deletion of parent record could cascade down, lookups could be cleared, and automation such as flows, triggers, and outbound messages might already have run on the loaded records, and therefore the plan includes the listing of automation to be deactivated during the load, the downstream effects that are irreversible, and the rollback goes through from children to parents in order not to leave orphans. In the third scenario, there is a partial transaction: the bulk operation commits in batches, hence a failure means that some batches are loaded and some are not, and the output files containing the list of successfully and unsuccessfully processed records are the authoritative list of what is there; the rollback set is calculated based on these output files and on the query on the legacy identifier, and these two are compared. And in the fourth scenario, the deletion semantics of the operation matter: record deleted through API goes to the Recycle Bin where it can be restored within a limited period, while hard delete goes around it. So the programme plan should clearly define what kind of deletion is used, and whether the recycle bin retention period is long enough compared to the decision period, since a soft-deleted record still retains its unique identifier value, which can prevent a reload.

The time required for the load is calculated in Section 3.15 and Table 10: at the assumed average of 40,000 records per minute the programme of 6 million records would take 150 minutes to load, two-pass design implies that the parents are loaded before the children, and the planned window is at least twice the pure load time, or about five hours for one load of all 6 million records. These figures are illustrative only and should be substituted by the measured throughput of the dress rehearsal.

4.5 Comparison with common practice

The common practices of the migrations rely on record counts, individual fields' format checks, and user acceptance testing near the end of the programme. They are necessary but not sufficient, because the literature cited above shows what these tools miss. Count checks cannot detect offsetting errors, format checks do not prove identity, and late user testing detects problems when they become too expensive to fix. The reliable programme design relies on the entity-level resolution with thresholds (Benjelloun *et* al., 2009; Papadakis *et* al., 2016), record-level audit with interval estimation (Brown *et* al., 2001; Wilson, 1927), and monitoring after cutover (Schelter *et* al., 2018). Table 13 compares the two approaches with regard to six aspects.

Table 13: Common migration practice versus the gate-based approach.

Feature	Common practice	Gate-based approach
Completeness check	Record counts by object	Counts plus control totals plus record-level audit with interval estimate
Identity	Duplicates removed by exact-match rules	Probabilistic matching with thresholds per business unit and a clerical band
Acceptance	User acceptance testing near go-live	Exit criteria set before the first wave and tested at every gate
Load control	Maximum parallelism for speed	Batch sizing for lock contention and ownership skew
Privacy	Production data in non-production systems	Masked test data; retention schedule for staging and error files
After go-live	Hypercare ticket queue	Scheduled quality indicators with named owners

The tool vendors and the system integrators could assist in this respect by making public the quantities used in this paper: a matching tool reporting the number of matches and unmatched records is hiding the information needed to set thresholds, whereas the tool reporting score distribution, the clerical band, and the effect of threshold adjustment is supporting the decision rule presented in Section 3.3. The loader reporting the error rows with the cause code allows to distinguish lock contention from data problems. Finally, the auditors could ask for the audit sample design, the interval estimate, and the decision against the criterion in order to have stronger evidence than a statement that the checks are performed (Elder *et* al., 2013; Harron *et* al., 2017).

4.6 Research agenda

The review raises issues related to practice. First, there are no empirical studies that would present criteria for acceptance of CRM migration validated against future operational results, thus Table 5 is based on propositions. Second, match thresholds are usually published without the labelled data used to determine them, therefore comparisons are not possible. Third, there is no measurement of costs associated with false matches in CRM settings, although they are the major reason why the asymmetric threshold should be used. Fourth, matcher that learn from examples have not been tested on multi-subsidiary data contaminated with realistic levels of noise (Li *et* al., 2021; Papadakis *et* al., 2016). Fifth, the interplay of data protection requirements and migration design, for instance, deletion of personal data right after its loading, has not been studied (Shastri *et* al., 2020). Table 14 presents these issues along with the minimum design that is needed to address them.

Table 14: Research agenda for CRM data migration validation.

No.	Question	Minimum design
1	Do proposed gate criteria predict post-migration data defects found by users?	Prospective study across several migrations with follow-up audits
2	How stable are match thresholds across business units and over time?	Repeated labelling studies with calibration curves
3	What is the cost of a false merge compared with a missed duplicate in CRM?	Case-based costing with service and sales data
4	Do learned matchers improve recall without raising false merges on multi-unit customer data?	Benchmark with held-out labelled pairs per unit
5	What is the cost of erasure and retention duties after load?	Performance and process measurement on target platforms
6	Which staging and rollback designs minimise the window of exposure?	Dress rehearsal comparisons

4.7 Programme-level evidence from enterprise system transitions

Relevance rule will govern this section: all papers below belong to adjacent domains and are included only where each of them proves one technical point of the framework. Chibwaye and Oteri (2023) and Oteri and Chibwaye (2025) relate to programme sequencing and post-cutover monitoring indicators of Section 3.16, respectively; Olagunju *et al.* (2024) - to capacity-based scheduling of waves; Agboola *et al.* (2022) and Ogeawuchi *et al.* (2022a, 2022b) – to load orchestration and pipeline automation in Gate 4; and Iwuanyanwu *et al.* (2025a, 2025b) – to the downstream analytics whose quality of recording is protected by Gates 1 to 4. The behavioural health review by Akinyelure (2025) has the lowest relevance and is retained only for the principle of governance according to which merged records have to be re-linked to the teams working with them, but it does not provide any migration controls. None of these sources provides defect rates in CRM migrations, none of them has been used for setting a quantitative criterion in Section 3, and none of them constitutes validation of the framework. The sources for the framework itself are CRM, entity-resolution and data quality literature of Sections 3.1 to 3.18 and where an article from this section contradicts them the latter prevails.

Studies of migrations of large enterprises system place data migration within a broader programme and support the gate-based approach used in this paper. A review of data-integrated ERP migration for large government workforce transformation includes topics of master data harmonisation, API-driven interoperability, middleware orchestration and federated data governance; and names data quality heterogeneity as one of the recurrent obstacles (Chibwaye & Oteri, 2023). This finding is consistent with the profiling and match-thresholds results presented above in terms of high defect concentration in some fields. A review of change management analytics for ERP migrations, cloud migrations and platform integrations includes topics of predictive analytics, process mining and adoption monitoring (Oteri & Chibwaye, 2025); and they are relevant to the post-cutover monitoring indicators proposed in this paper. A framework of performance optimisation for high-demand technology programmes includes demand and capacity modelling, constraint analysis and

KPI feedback loops (Olagunju *et* al., 2024); and the wave schedule and rehearsals proposed in this paper follow the same logic of sequencing by capacity. These sources are conceptual reviews and frameworks, thus they are informative regarding the design and do not include field measurements of migration defect rates.

Review of behavioural health data and public health decisions provides the parallel to migration governance. It states that data in clinical systems, online platforms and social services is increasing while policy remains unconnected to it (Akinyelure, 2025); and similarly migrated CRM is disconnected when merged records are not connected to the teams using them. The review concludes that contextual knowledge of the people represented by the data leads to more relevant and accurate interventions (Akinyelure, 2025); and this is consistent with the user-acceptance sampling and account-owner review in the design of audit presented here. Also it states that ethical data governance and organisational capacity are prerequisites for the use of this data (Akinyelure, 2025). The review is concerned with health policy and therefore it is included here only for the stated principles and not for migration measurements.

Practitioner reviews of cloud data engineering cover the context pipeline. A systematic review of cloud data migration techniques for platform integration in enterprise analytics (Agboola *et* al., 2022), a review of data orchestration and workflow automation (Ogeawuchi *et* al., 2022a), and a study of data pipeline automation for scalability (Ogeawuchi *et* al., 2022b) deal with load orchestration and automation, and thus relate to the load design of Gate 4. They cover analytics platforms, rather than CRM migration, and are listed in the references to illustrate their topics stated in the titles.

Research into enterprise data analytics covers the downstream consequences of the controls of migration that must be protected. A systematic review of AI based data analytics pipelines for large scale systems (Iwuanyanwu *et* al., 2025b) and a conceptual framework for enterprise data analytics governance and decision intelligence systems (Iwuanyanwu *et* al., 2025a) address pipeline reliability and governance, and these rely on the record quality established through Gates 1 to 4. They cover analytics pipelines in general, rather than CRM migration, and are listed in the references for the topics in their titles.

4.8 An evaluation strategy for the framework

The framework itself is provided as a synthesis, and should be subject to validation. A validation study would document, for each wave, the false-merge rate and the missed-duplicate rate, estimated from a labelled audit sample of merge decisions on the survivor map and of known duplicate pairs; the blocking recall, reported as the pairs completeness of Equation (9); the reconciliation discrepancy rate, defined as unexplained residuals per ten thousand records after the application of the survivor-map identity; the load failure rate, disaggregated by cause code from the job result files; and the post-cutover defect incidence, extracted from the indicators of Table 11. Table 15 summarises the measures, with their definitions, and the gate at which each is produced.

Table 15: Evaluation measures.

Measure	Definition	Data source	Gate
False-merge rate	Wrong merge count as per merge decisions made, based on an audited survivor file	Survivor map; adjudicated sample	3
Missed-duplicate rate	True duplicates after cutover with known duplicate pairs	Scheduled post-cutover match runs; labelled sample	3 and post-cutover
Blocking recall (PC)	True matching pairs captured in the candidate set, Equation (9)	Blocking audit on labelled pairs	3
Reconciliation discrepancy rate	Unexplained residuals per 10,000 records after the survivor-map identity	Gate 5 reconciliation reports	5
Load failure rate	Failed records per batch, by cause code	Job result files	4
Review workload	Share of candidate pairs directed to clerical review	Matcher score distribution	3
Post-cutover defect incidence	Defects found by users or by scheduled checks per month	Monitoring indicators of Table 11	Post-cutover

The threshold sensitivity analysis closes the validation strategy. Running the upper threshold through a range and recording, for each value, the resulting false-merge rate and the fraction of pairs going to clerical review generates the operating curve on which a group sets its position along the trade-off between merge risk and clerical effort. This should be repeated for each business unit, because the probabilities of Section 3.3 vary between units, and the resulting curves will indicate whether thresholds set on one unit carry over to others or need to be re-estimated, which is the re-estimation rule of the framework. Both the threshold sensitivity analysis and the measure plan above can be done on the waves of one migration and published without exposure of any organisational data, and thus without the confidentiality objection to validation results usually raised to block publication.

5. CONCLUSION

A successful large CRM migration hinges on the integrity of its data, and this is a quality that can be controlled and measured at certain points. The review has organised those points into five gates: profiling, cleansing, entity resolution, load, and reconciliation. Three weak points were identified in current practice and in the literature: validation focused on field formats and not on identity, match thresholds were tuned once but not re-estimated for each unit, and reconciliation used control totals blind to offsetting errors. The framework addresses each issue with a measurable control: a quality score for each mandatory field, a match weight with two thresholds and a clerical review band, a load plan adapted to platform limitations, and an auditable record-level check sized on the basis of its precision requirement with a correct interpretation of a clean sample. Used wave by wave and with named data owners, the gates provide a group with documentation for the decision to cut over, and give the single customer view a solid foundation.

The known findings that are the basis of the framework include the multidimensional view of data quality (Wang & Strong, 1996), the probabilistic decision rule of record linkage (Fellegi & Sunter, 1969), the sampling theory that determines the audit size (Cochran, 1977), the interval estimators that interpret it (Brown *et* al., 2001; Wilson, 1927), and the costs that have been associated with the deficiency of data quality (Redman, 1998; Haug *et* al., 2011). The five-gate process, the proposed criteria of its termination, the allocation and sampling size illustrations, and the accounting identity that follows from the survivor-map are the reviewer's contribution rather than the known findings. The major limitations of the work are in the fact that no migration of records has been so far investigated with the application of these gates. Therefore, the framework should be considered as a test design with the gate results registered and reported as per the specifications provided in Section 4.8.

REFERENCES

1. Abedjan, Z., Golab, L., & Naumann, F. (2015). Profiling relational data: A survey. *The VLDB Journal, 24*(4), 557-581. https://doi.org/10.1007/s00778-015-0389-y
2. Abraham, R., Schneider, J., & vom Brocke, J. (2019). Data governance: A conceptual framework, structured review, and research agenda. *International Journal of Information Management, 49*, 424-438. https://doi.org/10.1016/j.ijinfomgt.2019.07.008
3. Adesuyi, M. O., Akomolafe, O., Olaogun, B. O., Ndukwe, V. U., & Sakyi, J. K. (2023). Global interoperability model for distributed ledger-based cross-border payment systems. International Journal of Multidisciplinary Research and Growth Evaluation, 4(6), 1291-1300. https://doi.org/10.54660/.IJMRGE.2023.4.6.1291-1300
4. Afrihyia, E., Ojukwu, P. U., & Akinse, S. G. (2024). Privacy-preserving health data governance models: A comparative review of blockchain and cryptographic strategies in U.S. and developing healthcare systems. International Journal of Advanced Multidisciplinary Research and Studies, 4(6), 3071-3086. https://doi.org/10.62225/2583049X.2024.4.6.5919
5. Agboola, O. A., Owoade, S., Abayomi, A. A., Ogbuefi, E., Gbenle, T. P., & Alozie, C. E. (2022). A systematic review on cloud data migration methods and best practices for flawless integration of data in enterprise analytics platforms. *International Journal of Social Science Exceptional* Research.
6. Akinyelure, F. (2025). Using behavioral health data to innovate policy: Closing the feedback loop from community data insights into public health decision making. *International Journal of Innovative Science and Research Technology,* 10(7). https://doi.org/10.38124/ijisrt/25jul1532
7. Aliliele, C., Mbonu, I. S., Uzoka, E., & Iwuanyanwu, U. (2025). New directions in data lakehouse architecture design for enterprise data loss prevention and compliance management. Shodhshauryam, International Scientific Refereed Research Journal, 8(4), 171-213. https://doi.org/10.32628/SHISRRJ258474

8. Aneke, O. B., Sydney, E., & Adesemoye, A. C. (2023). Data integrity and governance in regulated life science operations: Reviewing data integrity failures, control measures and remediation. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 9(10), 538-568.
9. Annan, A. O. (2022). Data privacy governance models for cross-border digital platforms. International Journal of Multidisciplinary Research and Growth Evaluation, 3(6), 949-964.
10. Atakpa, M. I., & Abolaji, T. O. (2022). Privacy-preserving data architecture model for regulated industry analytics with GDPR and HIPAA compliance. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 8(3), 781-808. https://doi.org/10.32628/CSEIT22558
11. Azeez, L. O., & Badmus, O. (2024). An innovative data integration technique for improving accuracy and reliability of GHG inventory reporting. Global Multidisciplinary Perspectives Journal, 1(6), 166-181.
12. Azeroual, O., & Jha, M. (2021). Without data quality, there is no data migration. *Big Data and Cognitive Computing, 5*(2), Article 24. https://doi.org/10.3390/bdcc5020024
13. Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys, 41*(3), Article 16. https://doi.org/10.1145/1541880.1541883
14. Benjelloun, O., Garcia-Molina, H., Menestrina, D., Su, Q., Whang, S. E., & Widom, J. (2009). Swoosh: A generic approach to entity resolution. *The VLDB Journal, 18*(1), 255-276. https://doi.org/10.1007/s00778-008-0098-x
15. Boulding, W., Staelin, R., Ehret, M., & Johnston, W. J. (2005). Customer relationship management roadmap: What is known, potential pitfalls, and where to go. *Journal of Marketing,* 69(4), 155-166. https://doi.org/10.1509/jmkg.2005.69.4.155
16. Brodie, M. L., & Stonebraker, M. (1995). *Migrating legacy systems: Gateways, interfaces & the incremental approach*. Morgan Kaufmann.
17. Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science, 16*(2), 101-133. https://doi.org/10.1214/ss/1009213286
18. Cai, L., & Zhu, Y. (2015). Challenges of data quality and data quality assessment in the big data era. *Data Science Journal,* 14, Article 2. https://doi.org/10.5334/dsj-2015-002
19. Chen, I. J., & Popovich, K. (2003). Understanding customer relationship management (CRM): People, process and technology. *Business Process Management Journal, 9*(5), 672-688. https://doi.org/10.1108/14637150310496758
20. Chibwaye, I. P., & Oteri, O. (2023). Advancements and conceptual framework for data integrated ERP migration in large scale government workforce transformation. International Journal of Advanced Multidisciplinary Research and Studies, 3(6), 2712-2726.
21. Christen, P. (2012a). *Data Matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection.* Springer. Retrieved from https://link.springer.com/book/9783642311635
22. Christen, P. (2012b). A survey of indexing techniques for scalable record linkage and deduplication. *IEEE Transactions on Knowledge and Data Engineering, 24*(9), 1537-1555. https://doi.org/10.1109/TKDE.2011.127
23. Christophides, V., Efthymiou, V., Palpanas, T., Papadakis, G., & Stefanidis, K. (2021). An overview of end-to-end entity resolution for big data. *ACM Computing Surveys, 53*(6), Article 127. https://doi.org/10.1145/3418896
24. Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). Wiley.
25. Doan, A., Halevy, A., & Ives, Z. (2012). *Principles of data integration*. Morgan Kaufmann.
26. Dogbatsey, E. A., Ebhojie, O., & Oyeleye, A. O. (2019). Reconciliation control and financial workflow redesign in emerging market organizations: A conceptual framework. Zenodo. https://doi.org/10.5281/zenodo.20447103
27. Elder, R. J., Akresh, A. D., Glover, S. M., Higgs, J. L., & Liljegren, J. (2013). Audit sampling research: A synthesis and implications for future research. *Auditing: A Journal of Practice & Theory, 32*(Supplement 1), 99-129.
28. Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering, 19*(1), 1-16. https://doi.org/10.1109/TKDE.2007.250581
29. Enamorado, T., Fifield, B., & Imai, K. (2019). Using a probabilistic model to assist merging of large-scale administrative records. *American Political Science Review, 113*(2), 353-371. https://doi.org/10.1017/S0003055418000783
30. European Parliament & Council of the European Union. (2016). Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement

of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union, L 119*, 1-88. https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

31. Fan, W. (2015). Data quality: From theory to practice. *ACM SIGMOD Record, 44*(3), 7-18. https://doi.org/10.1145/2854006.2854008
32. Federal Republic of Nigeria. (2023). *Nigeria Data Protection Act, 2023*. Nigeria Data Protection Commission. https://ndpc.gov.ng/wp-content/uploads/2024/03/Nigeria_Data_Protection_Act_2023.pdf
33. Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association, 64*(328), 1183-1210. https://doi.org/10.1080/01621459.1969.10501049
34. Getoor, L., & Machanavajjhala, A. (2012). Entity resolution: Theory, practice & open challenges. *Proceedings of the VLDB Endowment, 5*(12), 2018-2019.
35. Hanley, J. A., & Lippman-Hand, A. (1983). If nothing goes wrong, is everything all right? Interpreting zero numerators. *JAMA, 249*(13), 1743-1745. https://doi.org/10.1001/jama.1983.03330370053031
36. Harron, K. L., Doidge, J. C., Knight, H. E., Gilbert, R. E., Goldstein, H., Cromwell, D. A., & van der Meulen, J. H. (2017). Guide to assessing linkage quality for analysis of linked data. *International Journal of Epidemiology,* 46(5), 1699-1710. https://doi.org/10.1093/ije/dyx177
37. Haug, A., Arlbjørn, J. S., & Pedersen, A. (2009). Classification model of ERP system data quality. *Industrial Management & Data Systems,* 109(8), 1053-1068. https://doi.org/10.1108/02635570910991292
38. Haug, A., Zachariassen, F., & van Liempd, D. (2011). Cost of poor data quality. *Journal of Industrial Engineering and Management,* 4(2), 168-193. https://doi.org/10.3926/jiem..v4n2.p168-193
39. Hernandez, M. A., & Stolfo, S. J. (1995). Merge/purge problem for large databases. *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data* (pp. 127-138). ACM. https://doi.org/10.1145/223784.223807
40. Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). *Data Quality and Record Linkage* Techniques. Springer. https://doi.org/10.1007/0-387-69505-2
41. Ilodigwe, L., & Adesemoye, A. C. (2021). Data backbone of health system transformation: Governance and architecture framework for interoperability, data quality and advanced analytics at the national scale. International Journal of Health and Pharmaceutical Research, 6(2), 52-76. https://doi.org/10.56201/ijhpr.vol.6.no2.2021.pg52.76
42. Ilyas, I. F., & Chu, X. (2015). Cleaning relational data: Trends in consistency and deduplication. *Foundations and Trends in Databases,* 5(4), 281-393. https://doi.org/10.1561/1900000045
43. ISO/IEC. (2008). *ISO/IEC 25012:2008 Software engineering: Software product Quality Requirements and Evaluation (SQuaRE): Data quality model*. International Organization for Standardization. https://www.iso.org/standard/35736.html
44. ISO/IEC. (2015). *ISO/IEC 25024:2015 Systems and software engineering: Systems and software Quality Requirements and Evaluation (SQuaRE): Measurement of data quality*. International Organization for Standardization. https://www.iso.org/standard/35749.html
45. Iwuanyanwu, C. C., Aliliele, C., & Seyi-Lande, O. B. (2025a). Conceptual framework for enterprise data analytics governance and decision intelligence systems. *Advances in IT and Digital Security,* 1(1), 49-66. https://journals.dacinsightpublishers.com/AIDY/article/view/23
46. Iwuanyanwu, C. C., Aliliele, C., & Seyi-Lande, O. B. (2025b). Systematic review of AI-based data analytics pipelines for large-scale systems. *Advances in IT and Digital Security,* 1(1), 1-16. https://journals.dacinsightpublishers.com/AIDY/article/view/26
47. Khatri, V., & Brown, C. V. (2010). Designing data governance. *Communications of the ACM, 53*(1), 148-152. https://doi.org/10.1145/1629175.1629210
48. Kim, W., Choi, B.-J., Hong, E.-K., Kim, S.-K., & Lee, D. (2003). A taxonomy of dirty data. *Data Mining and Knowledge Discovery, 7*(1), 81-99. https://doi.org/10.1023/A:1021564703268
49. Kimball, R., & Caserta, J. (2004). *The data warehouse ETL toolkit: Practical techniques for extracting, cleaning, conforming, and delivering data*. Wiley.
50. Krasnikov, A., Jayachandran, S., & Kumar, V. (2009). Impact of customer relationship management implementation on cost and profit efficiencies: Evidence from the US commercial banking industry. *Journal of Marketing,* 73(6), 61-76. https://doi.org/10.1509/jmkg.73.6.61
51. Köpcke, H., & Rahm, E. (2010). Frameworks for entity matching: A comparison. *Data & Knowledge Engineering, 69*(2), 197-210. https://doi.org/10.1016/j.datak.2009.10.003
52. Ladapo, O. O., Dosunmu, A. A., Jooda, D., & Abolaji, T. O. (2025). Migration of applications and information systems to cloud computing infrastructure: Insights from a south African retail bank.

International Journal of Multidisciplinary Research and Growth Evaluation, 6(6), 1361-1375. https://doi.org/10.54660/.IJMRGE.2025.6.6.1361-1375

53. Ladapo, O. O., Jooda, D., Dosunmu, A. A., & Abolaji, T. O. (2022). Digital transformation in practice: The best practices in migrating to cloud in enterprise context. Journal of Frontiers in Multidisciplinary Research, 3(1), 643-655. https://doi.org/10.54660/.JFMR.2022.3.1.643-655
54. Li, Y., Li, J., Suhara, Y., Doan, A., & Tan, W.-C. (2021). Deep entity matching with pre-trained language models. *Proceedings of the VLDB Endowment, 14*(1), 50-60. https://doi.org/10.14778/3421424.3421431
55. Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). L-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data, 1*(1), Article 3. https://doi.org/10.1145/1217299.1217302
56. Madnick, S. E., Wang, R. Y., Lee, Y. W., & Zhu, H. (2009). Overview and framework for data and information quality research. *ACM Journal of Data and Information Quality, 1*(1), Article 2. https://doi.org/10.1145/1515693.1516680
57. Maingot, M., & Quon, T. K. (2009). Sampling practices of internal auditors at corporations on the Standard & Poor's Toronto Stock Exchange Composite Index. *Accounting Perspectives, 8*(3), 215-234. https://doi.org/10.1506/ap.8.3.2
58. Mbonu, I. S., Aliliele, C., Iwuanyanwu, U., & Oluoha, O. M. (2018). Conceptual framework for legal and ethical risk modeling in enterprise data protection governance systems. Iconic Research and Engineering Journals, 2(2), 207-226. https://doi.org/10.64388/IREV2I2-1714911
59. Mbonu, I. S., Aliliele, C., Uzoka, E., & Oluoha, O. M. (2019). Review of comparative data protection regulations and secure cloud implementation strategies across jurisdictions. Iconic Research and Engineering Journals, 2(9), 482-501. https://doi.org/10.64388/IREV2I9-1714912
60. Mbonu, I. S., Iwuanyanwu, U., Aliliele, C., & Uzoka, E. (2022). Review of data protection impact assessment models in multi cloud financial infrastructure systems. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 8(1), 589-623. https://doi.org/10.32628/CSEIT25442
61. National Institute of Standards and Technology. (2020). *NIST Privacy Framework: A tool for improving privacy through enterprise risk management, version 1.0* (NIST Cybersecurity White Paper 10). U.S. Department of Commerce. https://doi.org/10.6028/NIST.CSWP.10
62. Naumann, F., & Herschel, M. (2010). *An introduction to duplicate detection*. Morgan & Claypool. https://doi.org/10.1007/978-3-031-01835-0
63. Ndupu, K. C., Anunagba, C. O., Nwakamma, S., & Oyesiji, S. O. (2026). Machine Learning for Biomolecular Data Wrangling and Curation: Review of Learned Extraction, Error Detection, Integration and Curation Workflow. Advances in IT and Digital Security, 1(6), 1-17.
64. Newcombe, H. B., Kennedy, J. M., Axford, S. J., & James, A. P. (1959). Automatic linkage of vital records. *Science, 130*(3381), 954-959. https://doi.org/10.1126/science.130.3381.954
65. Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: Literature review and classification. *Expert Systems with Applications,* 36(2), 2592-2602. https://doi.org/10.1016/j.eswa.2008.02.021
66. Nwangwu, C. V., Duvbiama-Owasanoye, O. A., & Akano, O. A. (2023). Technical Due Diligence and Supply Chain Integration in Energy M&A: Review of Large-Scale Asset Acquisitions. Shodhshauryam, International Scientific Refereed Research Journal, 6(2), 513-539.
67. Ogeawuchi, J. C., Uzoka, A. C., Alozie, C. E., Agboola, O. A., Gbenle, T. P., & Owoade, S. (2022a). Systematic Review of Data Orchestration and Workflow Automation in Modern Data Engineering for Scalable Business Intelligence. *International Journal of Social Science Exceptional* Research.
68. Ogeawuchi, J. C., Uzoka, A. C., Alozie, C. E., Agboola, O. A., Owoade, S., & Akpe, O. E. (2022b). Next-Generation Data Pipeline Automation for Enhancing Efficiency and Scalability in Business Intelligence Systems. *International Journal of Social Science Exceptional Research,* 1(1), 277-282.
69. Ogundairo, K. M. (2025). Hardening the autonomous value chain: Guard rails and multi-enterprise ETL architecture using Lean Six Sigma for agentic supply chain orchestration. IIARD International Journal of Economics and Business Management, 11(12), 441-478. https://doi.org/10.56201/ijebm.vol.11.no12.2025.pg441.478
70. Ogunwola, T. A., & Alozie, C. (2026). Enterprise firewall modernization for business continuity: A framework for secure migration and operational stability. Computer Science & IT Research Journal, 7(5), 347-361. https://doi.org/10.51594/csitrj.v7i5.2287

71. Okoruwa, P. O., Fadayomi, O., Akeju, B., Edivri, J., Ogbole, J. I., & Abolaji, T. O. (2020). Conceptual model for privacy-centric security engineering in digital and cloud computing systems. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 6(5), 371-389. https://doi.org/10.32628/IJSRCSEIT
72. Olagunju, E. O., Edivri, J., & Oteri, O. (2024). Applied performance optimization frameworks for managing high-demand enterprise technology programs and system rollouts. *International Journal of Advanced Multidisciplinary Research and Studies, 4*(6), 3059-3070. https://doi.org/10.62225/2583049X.2024.4.6.5915
73. Ominyi, M. (2020). Developing a conceptual framework for post-merger compliance integration in multi-jurisdictional banking transaction. IIARD International Journal of Banking and Finance Research, 6(3), 78-95. https://doi.org/10.56201/ijbfr.vol.6.no3.2020.pg78.95
74. Ominyi, M., & Anichukwueze, C. C. (2022). Conceptual framework for alignment of data protection compliance between GDPR, NDPR, and United States privacy legal frameworks. International Journal of Social Sciences and Management Research, 8(5), 148-174. https://doi.org/10.56201/ijssmr.v8.no5.2022.pg148.174
75. Orise, C., & Niniola, S. (2025). Data governance in digital health learning systems: Systematic review of frameworks and best practices. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 11(5), 520-538.
76. Osinem, H. E., Eribenne, F., Peteranaba, S. F., & Soyege, O. S. (2024). Cloud interoperability framework for secure exchange of health information between multi-organization systems. International Journal of Advanced Multidisciplinary Research and Studies, 4.
77. Oteri, O., & Chibwaye, I. P. (2025). Advances and conceptual models in enterprise level change management analytics for large scale organizational systems transitions. *Gyanshauryam, International Scientific Refereed Research Journal,* 8(5), 177-211.
78. Papadakis, G., Skoutas, D., Thanos, E., & Palpanas, T. (2020). Blocking and filtering techniques for entity resolution: A survey. *ACM Computing Surveys, 53*(2), Article 31. https://doi.org/10.1145/3377455
79. Papadakis, G., Svirsky, J., Gal, A., & Palpanas, T. (2016). Comparative analysis of approximate blocking techniques for entity resolution. *Proceedings of the VLDB Endowment, 9(9), 684-695. https://doi.org/10.14778/2947618.2947624*
80. Payne, A., & Frow, P. (2005). A strategic framework for customer relationship management. *Journal of Marketing, 69*(4), 167-176. https://doi.org/10.1509/jmkg.2005.69.4.167
81. Peteranaba, S. F., Osinem, H. E., Eribenne, F., & Soyege, O. S. (2022). Health information governance and regulatory compliance: Comparative frameworks for data verification, risk screening, and confidentiality in health and financial systems. International Journal of Science, Technology & Society, 6(1), 60-84.
82. Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM, 45*(4), 211-218. https://doi.org/10.1145/505248.506010
83. Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin, 23*(4), 3-13.
84. Redman, T. C. (1998). The impact of poor data quality on the typical enterprise. *Communications of the ACM, 41*(2), 79-82.
85. Reid, A., & Catterall, M. (2005). Invisible data quality issues in a CRM implementation. *Journal of Database Marketing & Customer Strategy Management, 12*(4), 305-314.
86. Rekatsinas, T., Chu, X., Ilyas, I. F., & Ré, C. (2017). HoloClean: Holistic data repairs with probabilistic inference. *Proceedings of the VLDB Endowment, 10(11), 1190-1201. https://doi.org/10.14778/3137628.3137631*
87. Salesforce. (n.d.-a). *Bulk API 2.0 and Bulk API developer guide*. Salesforce Developers. Retrieved October 22, 2026, from https://developer.salesforce.com/docs/platform/api-asynch/guide/bulk-api-2-0.html
88. Salesforce. (n.d.-b). *Designing record access for enterprise scale*. Salesforce Developers. Retrieved October 22, 2026, from https://developer.salesforce.com/docs/platform/draes/guide/draes-group-membership-data-skew.html
89. Sayers, A., Ben-Shlomo, Y., Blom, A. W., & Steele, F. (2016). Probabilistic record linkage. *International Journal of Epidemiology, 45*(3), 954-964. https://doi.org/10.1093/ije/dyv322
90. Schelter, S., Lange, D., Schmidt, P., Celikel, M., Biessmann, F., & Grafberger, A. (2018). Automating large-scale data quality verification. *Proceedings of the VLDB Endowment, 11*(12), 1781-1794. https://doi.org/10.14778/3229863.3229867

91. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems,* 28, 2503-2511.
92. Shastri, S., Banakar, V., Wasserman, M., Kumar, A., & Chidambaram, V. (2020). Understanding and benchmarking the impact of GDPR on database systems. *Proceedings of the VLDB Endowment, 13(7), 1064-1077. https://doi.org/10.14778/3384345.3384354*
93. Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10*(5), 557-570.
94. Tikkinen-Piri, C., Rohunen, A., & Markkula, J. (2018). EU General Data Protection Regulation: Changes and implications for personal data collecting companies. *Computer Law & Security Review, 34*(1), 134-153. https://doi.org/10.1016/j.clsr.2017.05.015
95. Uchechi, N., Ominyi, M., Anichukwueze, C. C., & Jones, B. C. (2023). Advancements towards standardizing un-standardized compliance controls for high volume non-traditional distribution channels. International Journal of Advanced Multidisciplinary Research and Studies, 3(6). https://doi.org/10.62225/2583049X.2023.3.6.6838
96. Vassiliadis, P. (2009). A survey on extract-transform-load technology. *International Journal of Data Warehousing and Mining,* 5(3), 1-27. https://doi.org/10.4018/jdwm.2009070101
97. Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5-33. https://doi.org/10.1080/07421222.1996.11518099
98. Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association, 22*(158), 209-212. https://doi.org/10.1080/01621459.1927.10502953
99. Xu, H., Nord, J. H., Brown, N., & Nord, G. D. (2002). Data quality issues in implementing an ERP. *Industrial Management & Data Systems, 102*(1), 47-58. https://doi.org/10.1108/02635570210414668
100. Zahari, N. M., Hussin, W. Y. W., Yussof, M. Y. M., & Saman, F. M. (2015). Data quality challenges during data migration. In *Soft computing in data science* (pp. 33-42). Springer. https://doi.org/10.1007/978-981-287-936-3_4