Certificate in Master Data Migration

Data Quality and Cleansing

Attribute Standardization
Related terms: Data Harmonization, Normalization
Ensures that attribute values follow a common format across all records. For example, dates may be stored as DD-MM-YYYY, MM-DD-YYYY, or ISO-8601; standardization selects one format and converts all entries. This improves matching, reporting, and downstream processing. Challenges include handling legacy systems with conflicting formats and preserving locale-specific nuances.

Attribute Mapping
Related terms: Data Mapping, Transformation Rules
Defines how source attributes correspond to target attributes in a master data model. For instance, a source field “Cust_ID” might map to the target “CustomerNumber”. Accurate mapping prevents data loss and duplication. Common pitfalls involve ambiguous field names and overlooked derived attributes.

Business Rules Validation
Related terms: Rule Engine, Data Governance
Applies logical conditions that reflect business policies, such as “A customer’s credit limit cannot exceed $100,000”. Validation flags records that violate these rules, allowing corrective action before migration. Complexity rises with inter-record dependencies and evolving regulations.

Data Cleansing
Related terms: Data Cleaning, Data Quality
The process of detecting and correcting inaccurate, incomplete, or inconsistent data. Techniques include de-duplication, standardization, and enrichment. Example: Correcting misspelled city names (“San Fransisco” to “San Francisco”). Challenges include large data volumes and balancing automation with manual review.

Data Consolidation
Related terms: Data Integration, Master Data Management
Merges data from multiple sources into a single, unified view. For instance, combining sales and support databases to create a comprehensive customer profile. Effective consolidation reduces redundancy but may reveal conflicting records that require resolution.

Data Governance
Related terms: Stewardship, Policy Framework
A set of policies, procedures, and responsibilities that ensure data is managed as a strategic asset. Governance establishes data ownership, quality standards, and compliance mechanisms. Implementing governance can be hindered by organizational silos and resistance to change.

Data Harmonization
Related terms: Attribute Standardization, Semantic Alignment

Aligns disparate data structures and vocabularies to a common reference model. Example: Reconciling “Product Category” codes from two ERP systems to a unified taxonomy. Harmonization requires deep domain knowledge and careful handling of legacy codes.

Data Integration
Related terms: ETL, Data Federation
Combines data from different origins, providing a coherent dataset for migration. Integration may involve real-time APIs or batch extracts. Pitfalls include mismatched data types and latency issues that affect data freshness.

Data Lineage
Related terms: Provenance, Traceability
Tracks the origin, movement, and transformation of data elements from source to target. Lineage diagrams help auditors verify that migration steps preserve data integrity. Maintaining accurate lineage can be difficult when legacy systems lack metadata.

Data Migration Strategy
Related terms: Lift-and-Shift, Phased Migration
Outlines the approach, timeline, and resources for moving master data. Strategies may be “big-bang” (all at once) or incremental. Choosing a strategy involves assessing risk, system downtime, and stakeholder readiness.

Data Quality Assessment
Related terms: DQ Scorecard, Profiling
Measures the fitness of data for its intended purpose using dimensions such as completeness, accuracy, and timeliness. Profiling tools generate statistics (e.G., % Of null values) that guide cleansing priorities. A challenge is establishing realistic thresholds that balance effort and benefit.

Data Quality Dimensions
Related terms: Accuracy, Consistency, Validity
Core attributes used to evaluate data health. Typical dimensions include completeness, uniqueness, timeliness, and conformity. For master data, uniqueness is critical to avoid duplicate customer records. Over-emphasis on one dimension can mask deficiencies in others.

Data Quality Metrics
Related terms: KPI, Scorecard
Quantitative measures derived from quality dimensions, such as “Duplicate Rate = 2.3%”. Metrics provide objective insight and enable monitoring over time. Selecting appropriate metrics requires alignment with business objectives and data governance policies.

Data Profiling
Related terms: Statistical Analysis, Data Discovery
Analyzes data to uncover patterns, anomalies, and distributions before cleansing. Profiling may reveal that a “PhoneNumber” field contains alphanumeric characters, indicating data entry errors. Profiling large datasets

demands efficient sampling techniques.

Data Stewardship
Related terms: Data Owner, Custodianship
Assigns responsibility for data quality, security, and lifecycle management. Stewards approve changes, resolve conflicts, and enforce standards. Effective stewardship often requires cross-functional collaboration and training.

Data Validation Rules
Related terms: Constraint Checks, Business Rules
Automated checks that ensure data meets predefined criteria before loading. Example: A rule that "PostalCode must be 5 digits". Validation can be performed at the source, during transformation, or in the target system. Over-restrictive rules may reject legitimate edge cases.

Data Warehouse
Related terms: OLAP, ETL
A central repository optimized for analytical queries. In master data migration, the warehouse may serve as a staging area for cleansing. Designing warehouses that accommodate evolving master data models poses scalability challenges.

De-duplication
Related terms: Record Matching, Survivorship
Identifies and merges duplicate records based on matching algorithms. For example, two customer rows with slightly different spellings of the same name may be consolidated. Choosing appropriate match thresholds and handling conflicting attribute values are common difficulties.

Entity Resolution
Related terms: Record Linkage, De-duplication
Extends de-duplication across disparate data sources to recognize that records refer to the same real-world entity. Techniques use deterministic keys (e.G., SSN) or probabilistic scoring. High-volume resolution can be computationally intensive.

ETL (Extract, Transform, Load)
Related terms: Data Integration, Data Pipeline
A three-step process to move data from source systems to a target repository. Extraction pulls raw data, transformation cleanses and reshapes it, and loading writes it to the destination. Designing robust ETL jobs requires handling error recovery and performance tuning.

Exact Match Rule
Related terms: Deterministic Matching, Key Constraint
A matching condition where two records are considered identical only if all specified fields match exactly. Useful for high-confidence merges, such as matching on a unique identifier. Limitations arise when data entry errors cause otherwise identical records to differ.

Extraction Logic
Related terms: Source Query, Data Pull
Defines how data is retrieved from source systems, often using SQL or API calls. Extraction must respect source performance constraints and security policies. Poorly designed extraction can cause system slowdowns or incomplete data capture.

Fuzzy Matching
Related terms: Approximate Matching, Levenshtein Distance
Compares records using similarity scores rather than exact equality, allowing for typographical errors or alternate spellings. Example: “Acme Corp” vs “Acme Corporation”. Tuning similarity thresholds is critical to balance false positives and false negatives.

Format Standardization
Related terms: Normalization, Pattern Enforcement
Converts data to a predefined format, such as enforcing uppercase for country codes (“us” → “US”). Standardization simplifies downstream processing and reporting. Edge cases include preserving meaningful case distinctions (e.G., “IPhone”).

Governance Framework
Related terms: Policy Suite, Compliance Model
A structured set of policies, procedures, and tools that guide data management activities. The framework defines roles, decision rights, and escalation paths for data quality issues. Implementing a framework often requires cultural change and executive sponsorship.

Hierarchical Data
Related terms: Parent-Child Relationships, Bill of Materials
Data organized in a tree-like structure, such as product categories or organizational charts. Cleansing hierarchical data may involve validating that each child has a valid parent and that loops are absent. Complex hierarchies can cause recursive processing challenges.

Identity Management
Related terms: Master Data, Unique Identifier
Controls the creation, maintenance, and deletion of unique identifiers for entities (e.G., CustomerID). Proper identity management prevents duplicate creation and supports referential integrity. Integration with authentication systems adds security considerations.

Informatica Data Quality
Related terms: Data Profiling Tool, Rule Designer
A commercial platform that provides profiling, cleansing, matching, and monitoring capabilities. It offers pre-built reference data sets for address validation, for instance. Licensing costs and steep learning curves can be barriers for smaller organizations.

Integration Testing
Related terms: System Test, End-to-End Test

Validates that migrated master data works correctly with downstream applications. Tests may include creating a new order using a migrated customer record. Test planning must cover edge cases and performance under realistic load.

ISO 8000
Related terms: Data Quality Standard, Data Governance
International standard that defines data quality criteria and measurement methods. It emphasizes accuracy, completeness, and timeliness. Adoption can provide a common language for stakeholders but may require extensive documentation to achieve compliance.

Key Attribute
Related terms: Primary Key, Business Key
An attribute (or combination) that uniquely identifies a record within a domain. For customers, a "CustomerNumber" often serves as the key. Selecting the correct key is essential to avoid duplicate creation and to support referential integrity.

Lookup Table
Related terms: Reference Data, Code Mapping
A static table that translates codes from source to target values (e.G., "NY" → "New York"). Lookups improve consistency and support downstream analytics. Maintaining up-to-date lookup tables can be a continual effort.

Master Data Management (MDM)
Related terms: Golden Record, Data Stewardship
A discipline and set of technologies for creating a single, authoritative source of master data across the enterprise. MDM includes governance, cleansing, and synchronization processes. Implementing MDM often involves complex integration with multiple legacy systems.

Metadata
Related terms: Data Dictionary, Schema
Data that describes other data, such as field definitions, data types, and lineage. Accurate metadata supports effective cleansing by clarifying expected formats and constraints. Legacy systems may have incomplete or outdated metadata, complicating migration.

Normalization
Related terms: Standardization, Denormalization
Transforms data to a common representation, such as converting all phone numbers to E.164 Format. Normalization reduces variability and aids matching. Over-normalization can strip culturally relevant nuances (e.G., Local dialing prefixes).

Obfuscation
Related terms: Data Masking, Privacy
Alters sensitive data (e.G., SSN) to protect privacy while preserving format for testing. Obfuscation must retain syntactic validity to avoid breaking downstream validation rules. Reversibility is a concern when the

original data is needed for production migration.

Operational Data Store (ODS)
Related terms: Staging Area, Transactional Data
A temporary repository that holds cleansed data before final loading into the target system. The ODS enables incremental loads and rollback if errors are detected. Managing storage and ensuring data freshness are common operational challenges.

Outlier Detection
Related terms: Anomaly Detection, Statistical Profiling
Identifies data points that deviate markedly from expected patterns, such as a purchase amount of $1,000,000 for a retail customer. Outliers may signal errors or genuine exceptions; analysts must decide whether to correct or retain them. Automated detection can generate false positives.

Pattern Matching
Related terms: Regex, Syntax Validation
Uses regular expressions to verify that data conforms to expected patterns (e.G., Email addresses). Pattern matching helps catch format errors early. Complex patterns can be difficult to maintain and may need updates as standards evolve.

Phased Migration
Related terms: Incremental Migration, Rollback Strategy
Executes migration in stages, often by business unit or geography, allowing issues to be addressed before full rollout. Phasing reduces risk but extends project timelines and requires careful coordination of dependencies.

Primary Key Constraint
Related terms: Uniqueness, Referential Integrity
A database rule that ensures each record’s key value is unique and not null. Violations during load indicate duplicate or missing identifiers, prompting cleansing actions. Enforcing constraints early can prevent downstream data corruption.

Profiling Report
Related terms: Data Quality Assessment, Statistics Summary
A document generated by profiling tools that outlines data characteristics (e.G., Distinct count, null rate). Reports guide prioritization of cleansing tasks. Interpreting large reports requires domain expertise to distinguish critical issues from noise.

Quality Scorecard
Related terms: KPI Dashboard, Data Quality Metrics
Visual representation of data quality performance across dimensions, often using traffic-light indicators. Scorecards enable stakeholders to track improvement over time. Over-reliance on a single metric can mask underlying problems.

Reference Data
Related terms: Lookup Table, Code List
Static data that provides context for transactional data, such as country codes or product categories. Maintaining accurate reference data is essential for consistent cleansing. Changes in reference standards (e.G., ISO country updates) require coordinated updates.

Regulatory Compliance
Related terms: GDPR, HIPAA
Ensures that data handling meets legal requirements for privacy, security, and reporting. Compliance checks may mandate removal of personally identifiable information before migration. Balancing compliance with data utility can be challenging.

Record Linking
Related terms: Entity Resolution, Fuzzy Matching
Connects records that refer to the same entity across systems, often using a combination of deterministic and probabilistic methods. Effective linking reduces duplication and improves data completeness. High-volume linking demands scalable algorithms.

Reference Data Management
Related terms: Master Data Management, Lookup Maintenance
Processes for governing, updating, and distributing reference data sets. Centralized management prevents divergent code lists across applications. Governance must address versioning and backward compatibility.

Reconciliation
Related terms: Data Validation, Audit Trail
Compares source and target data to confirm that migration preserved record counts, totals, and relationships. Reconciliation may include row-level checks (e.G., Checksum) and aggregate checks (e.G., Total sales). Discrepancies often trace back to cleansing rules that altered values.

Redundancy Elimination
Related terms: De-duplication, Consolidation
Removes duplicate copies of master records, ensuring a single source of truth. This step improves performance and reduces storage costs. Care must be taken to preserve historic relationships (e.G., Orders linked to the duplicate record).

Reference Model
Related terms: Canonical Model, Data Architecture
A standard representation of entities and relationships used as a blueprint for migration. Aligning source schemas to the reference model facilitates mapping and validation. Divergence from the model can cause mapping gaps.

Reliability
Related terms: Data Consistency, Availability
Measures the extent to which data can be trusted over time. Reliable master data supports downstream

analytics and operational processes. Reliability may be compromised by intermittent source system outages during extraction.

Rollback Plan
Related terms: Contingency Plan, Recovery Strategy
Defines steps to revert the target system to its pre-migration state if critical errors occur. A robust plan includes backup procedures, data versioning, and clear decision criteria. Inadequate rollback planning can extend downtime and increase risk.

Schema Mapping
Related terms: Data Mapping, Structural Transformation
Describes how source tables, fields, and relationships correspond to target structures. For example, mapping a flat “Address” table into a normalized “Customer” and “Location” hierarchy. Complex schemas may require intermediate staging tables.

Segmentation
Related terms: Data Partitioning, Target Grouping
Divides master data into logical groups (e.G., By region) to facilitate phased migration or targeted cleansing. Segmentation helps allocate resources and monitor progress. Inconsistent segment definitions can lead to overlap or gaps.

Standard Operating Procedure (SOP)
Related terms: Process Documentation, Best Practices
Written instructions that define how cleansing activities should be performed. SOPs promote consistency across team members and support auditability. Keeping SOPs up-to-date as tools evolve is an ongoing effort.

Survivorship Rules
Related terms: Merge Logic, Attribute Prioritization
Determine which attribute value to retain when duplicate records are merged (e.G., Latest address vs. Most complete profile). Rules may be based on source system trust, timestamp, or data quality score. Poorly defined rules can result in loss of critical information.

Synonym Management
Related terms: Terminology Alignment, Semantic Mapping
Handles multiple terms that refer to the same concept, such as “Dept.” And “Department”. Managing synonyms improves searchability and matching accuracy. Maintaining synonym lists requires ongoing linguistic review.

System of Record (SOR)
Related terms: Source of Truth, Master Data Source
The authoritative system that holds the definitive version of a data element. Identifying the SOR is crucial for conflict resolution during migration. Multiple SORs for the same entity can cause data disputes.

Target Data Model
Related terms: Canonical Model, Schema Design
The structured representation of master data in the destination environment. The model defines entities, attributes, relationships, and constraints. Designing a flexible target model supports future enhancements but may increase initial complexity.

Technical Validation
Related terms: Schema Validation, Data Type Checks
Ensures that data conforms to technical specifications such as field lengths, data types, and encoding. For example, verifying that a numeric field does not contain alphabetic characters. Technical validation is often automated but must be complemented by business validation.

Temporal Data
Related terms: Effective Dates, Historical Tracking
Data that changes over time, such as price lists or address histories. Cleansing temporal data requires preserving change history while ensuring current records are accurate. Handling overlapping effective dates can be tricky.

Transformation Logic
Related terms: ETL, Business Rules
The set of operations that convert source data into the target format, including calculations, concatenations, and lookups. Well-documented logic aids debugging and auditability. Overly complex transformations increase maintenance burden.

Unique Constraint
Related terms: Primary Key, Duplicate Prevention
A database rule that ensures a combination of fields remains unique across records. Violations signal data quality issues that must be resolved before load. Implementing constraints early helps catch errors early in the pipeline.

Validation Framework
Related terms: Testing Suite, Quality Gates
A structured set of tests and checks applied at each migration stage to verify data integrity. Frameworks may include unit tests for transformations, integration tests for end-to-end flow, and performance tests. Maintaining the framework as requirements evolve can be resource-intensive.

Value Mapping
Related terms: Lookup Table, Code Translation
Translates source values to target equivalents, such as mapping “Y”/“N” to “Yes”/“No”. Accurate value mapping prevents misinterpretation in downstream systems. Inconsistent source values (e.G., “Yes”, “Y”, “1”) increase mapping complexity.

Verification
Related terms: Validation, Audit

The process of confirming that data has been correctly migrated and meets quality standards. Verification may involve spot checks, automated scripts, and stakeholder sign-off. Insufficient verification can leave hidden defects that surface later.

Workflow Automation
Related terms: Orchestration, Process Engine
Uses tools to automate repetitive cleansing steps, such as rule application or notification of data stewards. Automation speeds up migration and reduces manual error. Designing flexible workflows that accommodate exception handling is essential.

XML Data Exchange
Related terms: EDI, Web Services
A format for transmitting structured data between systems. XML schemas define element names, data types, and constraints. Cleansing XML data may involve validating against XSDs and normalizing namespaces. Large XML files can cause performance bottlenecks.

Zero-Day Data
Related terms: Real-time Data, Streaming
Data generated or updated on the day of migration, requiring immediate capture to avoid loss. Incorporating zero-day data often demands incremental loads or change data capture mechanisms. Ensuring consistency between batch and real-time streams poses synchronization challenges.