

Data Cleansing and Transformation Rules

Document Sample

1. Document Overview

This document outlines standardized data cleansing and transformation rules for use in all ETL processes. The purpose is to ensure data consistency, accuracy, and readiness for analytics and reporting.

2. Data Fields and Cleansing Rules

Field Name	Rule Description	Example Input	Transformed Output
Email Address	- Trim leading/trailing spaces - Convert to lowercase - Remove invalid characters	" John.Doe@Example.COM "	john.doe@example.com
Phone Number	- Remove non-numeric characters - Standardize to international format	(800) 123-4567	+18001234567
Date of Birth	- Parse multiple date formats - Output as YYYY-MM-DD - Validate date is in the past	3/12/90	1990-03-12
Name	- Trim spaces - Capitalize first letters - Remove numeric/invalid chars	aNNA o'cONnOr123	Anna O'Connor
Address	- Remove duplicate whitespace - Standardize abbreviations (St., Ave., etc.) - Title case conversion	22 main st.	22 Main St.

3. General Cleansing Rules

- Remove records with required fields missing.
- Replace nulls in non-critical fields with default values (if applicable).
- Trim all string fields.
- Remove leading and trailing whitespace for all fields.
- Remove duplicate records based on unique key.

4. Transformation Rules

- Standardize date formats to ISO 8601 across all date fields.
- Map categorical fields to defined set of allowed values.
- Split full name into first and last name using space as separator.
- Calculate new fields when needed (e.g., age from date of birth).
- Aggregate data at the required granularity (daily, monthly, etc.).

5. Reference Data and Mapping Tables

Source Value	Mapped Value
NY, N.Y., New York	New York
CA, Calif, California	California
USA, United States, US	United States

6. Exceptions and Data Quality Rules

- Flag records missing critical fields (e.g., customer ID, email).
- Log all records failing cleansing or transformation for review.
- Set up validations for field ranges (e.g., age between 0 and 120).
- Review and update rules quarterly.

7. Document Revision History

Date	Version	Changes	Author
2024-06-05	1.0	Initial sample document	Data Team