Job Sample – Part 1 Report (Data Cleaning & Organization in Salesforce)

Candidate: Arath Mendivil Mora

Date: September 16, 2025

Dataset & Loading

File: salesforce report.xlsx

Issues Found

- Inconsistent casing and extra whitespace in text fields.
- Missing values across numeric and text columns.
- Potentially malformed emails (format) and phones (length/characters).
- Date-like columns with mixed formats/timezones.
- Duplicate records (by ID or full row).

Cleaning Steps

- Trim & normalization of text columns (emails to lowercase; cities in Title Case).
- Validation & standardization for emails (regex) and phones (E.164-like length check).
- Date parsing using flexible patterns; conversion to consistent datetime type.
- Numeric coercion for amount-like fields; median imputation where partially missing.
- Deduplication (by Id when present; otherwise full-row), keeping the first occurrence.

API Integration (Public)

- Source: Open-Meteo (no API key).
- Enrichment: Daily mean temperature (°C) and precipitation (mm).
- Join keys: (City, Date).
- Output: dataset_enriched.csv.

Deliverables

- dataset_clean.csv — cleaned dataset.
- dataset_enriched.csv — cleaned + API-enriched dataset.
- Part1_Report.md / .docx — summary of issues, steps and recommendations.

Proposed Data Structuring / CRM (Salesforce)

- Account: AccountId (PK), Name, Industry, BillingCountry, CreatedDate
- Contact: ContactId (PK), AccountId (FK), FirstName, LastName, Email (unique), Phone (E.164)
- Lead: LeadId (PK), Company, Email, Phone, LeadSource, Status
- Interaction__c: InteractionId (PK), AccountId/ContactId (FK), InteractionDate, Type, Amount__c, SourceSystem__c, CreatedTS__c
- Location__c: LocationId (PK), City, Country, Latitude, Longitude

- WeatherDaily__c: WeatherId (PK), LocationId (FK), Date, TempMeanC__c, PrecipMM__c, Source="Open-Meteo"

Recommended Improvements

- Field validations & rules: enforce email regex, E.164 phone format, required fields; add date sanity checks.
- Unique keys & dedup logic: define canonical PKs and merge rules; automate deduplication on ingestion.
- Standardized time & locale: store dates in UTC; normalize country/city names using a reference table.
- Data lineage & monitoring: track SourceSystem, IngestedAt, and row hash; monitor quality KPIs.
- Incremental ETL: schedule periodic loads with idempotent upserts; keep backfill scripts for corrections.
- Location normalization: maintain a Location dimension and use IDs in fact tables.
- API enrichment governance: persist enrichment tables with source and refresh cadence; log retries/errors.
- Documentation & access: data dictionary and access policy aligned with CRM roles.