**Phase 1 – Raw Data Exploration**

**Goal:** Understand each dataset’s structure and relationships.

1. **List All Datasets**
   * Write down the dataset names and a short description of what each contains.
2. **Check Column Details**
   * For each dataset:
     + List column names.
     + Identify their data types (string, number, date, etc.).
     + Write the purpose of each column.
3. **Identify Key Columns**
   * Determine which columns can link datasets (PK/FK relationships).
   * Example: customer\_id, order\_id.
4. **Relationship Mapping**
   * Draw a quick diagram showing how datasets connect.
   * Note one-to-many or many-to-many relationships.

**Phase 2 – Data Quality Checks**

**Goal:** Find issues without modifying the raw data.

1. **Missing Values**
   * Write SQL queries to count NULL/empty values in each column.
2. **Duplicate Records**
   * Write SQL queries to detect duplicates based on unique identifiers.
3. **Format Inconsistencies**
   * Check date formats.
   * Check text casing.
   * Check numeric precision.
4. **Orphan Records**
   * Verify if foreign key references exist in parent tables.
5. **Documentation**
   * Record issues found in a table format (Issue → Table → Column → Count → Notes).

**Phase 3 – Master Table Design**

**Goal:** Plan the final integrated table.

1. **Select Essential Columns**
   * Decide which columns from each dataset go into the Master Table.
2. **Define PKs & FKs**
   * Assign primary keys and foreign keys.
3. **Data Types & Rules**
   * Set column data types.
   * Define constraints (e.g., NOT NULL, UNIQUE).
4. **Schema Draft**
   * Create a table schema document with:
     + Column name
     + Data type
     + Constraints
     + Source dataset

**Phase 4 – ETL Procedure**

**Goal:** Extract, transform, and load data into the Master Table.

1. **Extract**
   * Write SELECT queries to fetch only relevant columns from source datasets.
2. **Transform**
   * Apply cleaning rules:
     + Remove duplicates.
     + Standardize formats.
     + Handle missing values (decide: fill, ignore, or flag).
3. **Load**
   * Insert cleaned data into the Master Table.
   * Maintain relationships.
4. **Automation**
   * Create a stored procedure or script for repeatable ETL execution.

**Phase 5 – Validation & Refinement**

**Goal:** Ensure the Master Table is correct.

1. **Record Count Validation**
   * Compare record counts pre- and post-load.
2. **Duplicate Check**
   * Confirm no duplicates exist.
3. **Missing Data Check**
   * Ensure critical columns have no missing values.
4. **Foreign Key Integrity**
   * Validate all FK references exist.
5. **Data Type Check**
   * Ensure all columns match their intended data types.
6. **ETL Adjustment**
   * If issues found, fix ETL script and reload.

**Phase 6 – Final Documentation**

**Goal:** Create a Data Quality Report.

1. **Summary**
   * What datasets were used.
   * Structure of the Master Table.
2. **Issues Found**
   * List all data quality issues and their counts.
3. **Solutions Applied**
   * How each issue was handled in ETL.
4. **Validation Results**
   * Record counts.
   * Duplicate status.
   * Missing value status.
   * Format consistency.