1. **Components of the Architecture**

**Key Components:**

1. **Input Data**:
   * **Source**: Excel file Online Retail.xlsx.
   * **Description**: Raw dataset containing transaction records.
2. **Data Cleaning and Standardization**:
   * **Remove Duplicate**: Identifies and eliminates duplicate records based on specific columns.
   * **Outlier Detection**: Uses the IQR method to flag and replace outliers.
   * **Missing Value Handling**: Fills or drops missing values for critical columns.
3. **Feature Engineering**:
   * Adds a TransactionType column to distinguish sales from returns.
   * Reorders and updates columns for consistency.
4. **Database Storage**:
   * SQLite database to store and query cleaned data for analysis.
5. **Data Normalization**:
   * Applies Min-Max scaling and Z-score normalization for numeric columns.
   * Standardizes date columns to ISO 8601 format.
6. **Output**:
   * Cleaned data saved to Online\_retail\_unique\_updated.xlsx.
   * Normalized and processed data available for analysis or modeling.

**II) Determine the Data Flow**

**Data Flow:**

1. **Input**:
   * Load raw data from Online Retail.xlsx.
2. **Data Cleaning**:
   * Check for and remove duplicates.
   * Detect and handle outliers.
   * Fill or handle missing values.
3. **Feature Engineering**:
   * Add calculated columns (e.g., TransactionType).
   * Standardize and reorder columns.
4. **Normalization**:
   * Normalize numerical data using Min-Max scaling and Z-score normalization.
   * Format dates to ISO format.
5. **Database Storage**:
   * Load cleaned data into an SQLite database for efficient querying.
6. **Output**:
   * Save cleaned data to a new Excel file (Online\_retail\_unique\_updated.xlsx).

A diagram of a system

Description automatically generated