This documentation presents a comprehensive analysis and remediation plan for an IMDb dataset. It addresses inconsistencies, missing data, and structural anomalies within various tables, JSON formats, and TSV datasets. By identifying these issues and proposing appropriate solutions, this document aims to streamline the dataset for enhanced usability and reliability.

Covering a range of challenges such as inconsistent date formats, missing values in crucial columns, comma-separated entries, and special characters in timestamps, each section outlines specific problems and their corresponding resolutions.

**Database: imdb\_name\_basics**

1. **Column: birthYear**
   * **Problem:** Some entries contain "\N," indicating missing or unknown birth years.
   * **Proposed Solution:** Replace "\N" with a suitable representation for missing values, such as converting it to NULL or using a placeholder for unknown birth years.
2. **Column: deathYear**
   * **Problem:** Some entries contain "\N," indicating missing or unknown death years.
   * **Proposed Solution:** Replace "\N" with an appropriate representation for missing values, like converting it to NULL or using a placeholder for unknown death years.
3. **Column: knownForTitles**
   * **Problem:** Contains comma-separated entries listing titles individuals are known for.
   * **Proposed Solution:** Normalize the data by splitting these comma-separated entries into separate rows. Each row should represent a single title associated with an individual, avoiding the storage of multiple values in a single cell.
4. **Column: DI\_Create\_DT**
   * **Problem:** Timestamps are in a specific format ("YYYY-MM-DD HH:MM:SS").
   * **Proposed Solution:** Ensure consistency in date formats by converting timestamps to a standardized format (e.g., "YYYY-MM-DD HH:MM:SS").

**Database: imdb\_title\_akas**

1. **Column: language**
   * **Problem:** The 'language' column contains numerous entries marked as "\N," indicating missing or unknown language data for specific titles.
   * **Proposed Solution:** Replace "\N" entries in the 'language' column with an appropriate representation for missing values, such as NULL or a placeholder value. This will enhance the data's integrity and usability, allowing for more accurate language-based analysis or categorization of titles.
2. **Column: attributes**
   * **Problem:** The 'attributes' column contains entries marked as "\N," indicating missing or unknown attribute data for certain titles.
   * **Proposed Solution:** Replace "\N" entries in the 'attributes' column with an appropriate representation for missing values, such as NULL or a placeholder value. By addressing these missing attribute values, the dataset's completeness and accuracy can be improved, ensuring better analysis and understanding of title attributes.

**Database: imdb\_title\_basics**

1. **Column: startYear**
   * **Problem:** Contains the starting year of the title's release.
   * **Proposed Solution:** Handle missing or unknown values (e.g., "\N") by converting them to NULL or other suitable representations for missing data.
2. **Column: endYear**
   * **Problem:** Indicates the ending year for TV shows (movies will have "\N").
   * **Proposed Solution:** Ensure uniformity in representing missing or unknown values (e.g., convert "\N" to NULL).
3. **Column: runtimeMinutes**
   * **Problem:** Indicates the runtime of titles in minutes.
   * **Proposed Solution:** Handle missing or unknown values (e.g., "\N") by converting them to NULL or other suitable representations for missing data.
4. **Column: genres**
   * **Problem:** Contains genres associated with each title as strings separated by commas.
   * **Proposed Solution:** Normalize the genres data by splitting the entries into separate rows to avoid storing multiple values in a single cell.
5. **Column: DI\_Create\_DT**
   * **Problem:** Timestamps are in a specific format ("YYYY-MM-DD HH:MM:SS").
   * **Proposed Solution:** Ensure consistency in date formats within the 'DI\_Create\_DT' column. Validate and standardize all timestamps to a consistent format, such as "YYYY-MM-DD HH:MM:SS". This action ensures uniformity in storing and handling timestamp data across the dataset.

**Database: imdb\_title\_crew**

1. **Multiple Directors per Entry**
   * **Affected Data:** 'directors'
   * **Problem Description:** Multiple director IDs are listed in the same field, separated by commas.
   * **Proposed Solution:** Normalize the 'directors' data by splitting the IDs into separate rows or columns for individual directors per title entry. Ensure uniform representation of missing or unknown director IDs (e.g., converting "\N" to NULL).
2. **Writers Data in Comma-Separated Entries**
   * **Affected Data:** 'writers'
   * **Problem Description:** Writers' IDs are listed in a comma-separated format.
   * **Proposed Solution:** Normalize the 'writers' data by splitting the IDs into separate rows or columns for individual writers per title entry. Handle missing or unknown writer IDs (e.g., converting "\N" to NULL) for consistent representation.
3. **Handling Special Characters in 'DI\_Create\_DT'**
   * **Affected Data:** 'DI\_Create\_DT'
   * **Problem Description:** Timestamps may have special characters or non-uniform format.
   * **Proposed Solution:** Standardize the 'DI\_Create\_DT' column by removing any special characters or ensuring a consistent timestamp format (e.g., "YYYY-MM-DD HH:MM:SS"). Convert the 'DI\_Create\_DT' column to an appropriate timestamp data type for consistency.

**Database: imdb\_title\_ratings**

1. **Column: DI\_Create\_DT**
   * **Problem:** Timestamps are in a specific format ("YYYY-MM-DD HH:MM:SS").
   * **Proposed Solution:** Ensure consistency in date formats by converting timestamps to a standardized format (e.g., "YYYY-MM-DD HH:MM:SS").

**JSON: title\_basics**

1. **Inconsistent Genre Representation**
   * **Affected Data:** All entries
   * **Problem Description:** Genres are presented as strings separated by commas.
   * **Proposed Solution:** Normalize the 'genres' data to enhance clarity and ease of analysis. Split 'genres' into separate rows to ensure each genre occupies an individual record, eliminating string separation issues.
2. **Year Representation**
   * **Affected Data:** All entries
   * **Problem Description:** Years are represented as strings.
   * **Proposed Solution:** Standardize 'startYear' and 'endYear' columns by converting them into appropriate date data types (year). Transform these columns to a date-based format for improved analysis, sorting, and search capabilities.
3. **Cleaning Special Characters**
   * **Affected Data:** 'runtimeMinutes'
   * **Problem Description:** Runtime is represented as a string.
   * **Proposed Solution:** Cleanse 'runtimeMinutes' by removing any special characters or non-numeric symbols. Convert the 'runtimeMinutes' column to a numerical data type to ensure consistency and facilitate accurate runtime analysis.

**JSON: name\_basics**

1. **Inconsistent Naming Format**
   * **Affected Data:** 'primaryName'
   * **Problem Description:** Names contain additional aliases in parentheses.
   * **Proposed Solution:** Normalize 'primaryName' by removing aliases or alternate names from the main name. Store aliases in a separate column if needed for reference.
2. **Representation of Birth and Death Year**
   * **Affected Data:** 'birthYear', 'deathYear'
   * **Problem Description:** Years are represented as strings with a special character ('\N' for missing data).
   * **Proposed Solution:** Convert 'birthYear' and 'deathYear' into appropriate date data types (year) for consistent data representation. Replace '\N' values with NULL or another suitable representation for missing or unknown data.
3. **Handling Professional Roles**
   * **Affected Data:** 'primaryProfession'
   * **Problem Description:** Multiple professions are listed in a single string separated by commas.
   * **Proposed Solution:** Normalize 'primaryProfession' by splitting them into separate rows or columns to represent each profession individually.
4. **Management of Known Titles**
   * **Affected Data:** 'knownForTitles'
   * **Problem Description:** Titles are stored as a string with identifiers separated by commas.
   * **Proposed Solution:** Normalize 'knownForTitles' by splitting them into separate rows or columns to represent each title individually. Create a mapping table to link titles with other dataset tables if needed.

**TSV: Dataset**

1. **Date Representation**
   * **Affected Column:** 'Date'
   * **Problem Description:** Dates are represented in the format "Month Day, Year".
   * **Proposed Solution:** Convert the 'Date' column to a standard date format (e.g., YYYY-MM-DD) for consistency and better analysis. Ensure uniform representation of dates across the dataset.
2. **Handling Percentage Values**
   * **Affected Columns:** '%YD', '%LW'
   * **Problem Description:** Columns contain percentage values with percentage symbols.
   * **Proposed Solution:** Remove the percentage symbols from '%YD' and '%LW' columns. Convert the columns to a numerical data type for ease of analysis.
3. **Monetary Values in Gross Columns**
   * **Affected Columns:** 'Gross', 'Per Theater', 'Total Gross'
   * **Problem Description:** Monetary values are represented as strings with dollar symbols.
   * **Proposed Solution:** Remove the dollar symbols from 'Gross', 'Per Theater', and 'Total Gross' columns. Convert the columns to a numerical data type for consistent monetary representation.