MLDS Center

Maryland Longitudinal Data System

Better Data • Informed Choices • Improved Results

Identity Matching in the MLDS

July 23, 2024

I. Summary

This document details the process for managing data in the Master Data Management (MDM) component of the Maryland Longitudinal Data System (MLDS). Data management in the MDM consists of a series of steps, from parsing of the source data file, through identity matching, to loading of data to the Operational Data Store (ODS).

Accurately linking data over time and across sources is critical in a longitudinal data system. Identity matching encapsulates the process of determining the unique individual for a given record from a source file. The methods for matching identities within the MLDS to various data sets must be fully defined and comprehensive.

The purpose of this document is to explain the MLDS Center's matching algorithms and how they are used as part of the identity process within the Master Data Management database. Documentation on identities moving from MDM to ODS will be provided in other process documentation.

Overview

The MLDS contains data from over 35 different education and workforce collections across multiple sources, with data spanning from 2008 to present-day. There are currently about 3.9 million identities in the MLDS. The matching and identity matching between these sources is essential to the accuracy and quality of the data in the system.

II. Definition of Deterministic and Probabilistic

There are two matching methods the MLDS Center utilizes: Deterministic and Probabilistic. Both methods are used when agency collection source data are matched with Motor Vehicle Administration (MVA) data AND when agency collection source data are matched with MDM data. It is important to understand the definitions of deterministic and probabilistic in order to fully comprehend the steps in the matching process.

Deterministic

The Center defines a match as deterministic when the values match exactly between the agency collection source and the Master Data Management (MDM) database (shown later in Tables 1 and 3). MLDS data analysts conduct deterministic matching through a specific order (detailed below beginning in Step 4), based on what personally identifiable information (PII) is available from source files.

Probabilistic

The Center defines probabilistic matching as an approach to measure the probability that two records are for the same individual. Tables 2 and 4 below in Steps 4 and 5 show the order for probabilistic matching.

The Center uses the **Vladimir Levenshtein Method** (VLM) which is one of the industry standards in probabilistic matching. The VLM method identifies near matches of persons with similar spellings of

Names, similar Dates of Birth and potentially transposed SSNs within the agency data collection source file.

The "Edit Distance" or "Levenshtein Distance" measures the similarity between two strings (s1 and s2) by counting the number of character changes (e.g. insertions, updates) required to transform the first string into the second. The distance is the number of insertions, deletions, or substitutions required to transform s1 to s2. See the following example below:

Distance of 0:

```
select utl_match.edit_distance('Ripken','Ripken') edist from
dual;
-- Result:  0
```

Distance of 1:

```
select utl_match.edit_distance('Mathew','Matthew') edist from
dual;
-- Result:  1
```

Distance of 2:

```
select utl_match.edit_distance('Shawn','Sean') edist from dual;
-- Result:  2
```

The Center uses the Oracle 'utl_match.edit_distance' utility and in some cases Jaro-Winkler algorithm to perform probabilistic matching and calculate the edit distance between strings. The utility enables the Center to identify near matches of persons with a very high level of confidence. The recommended tolerance for edit distance is between 0 and 2, which is the threshold the MLDS Center follows, and is considered a high confidence match.

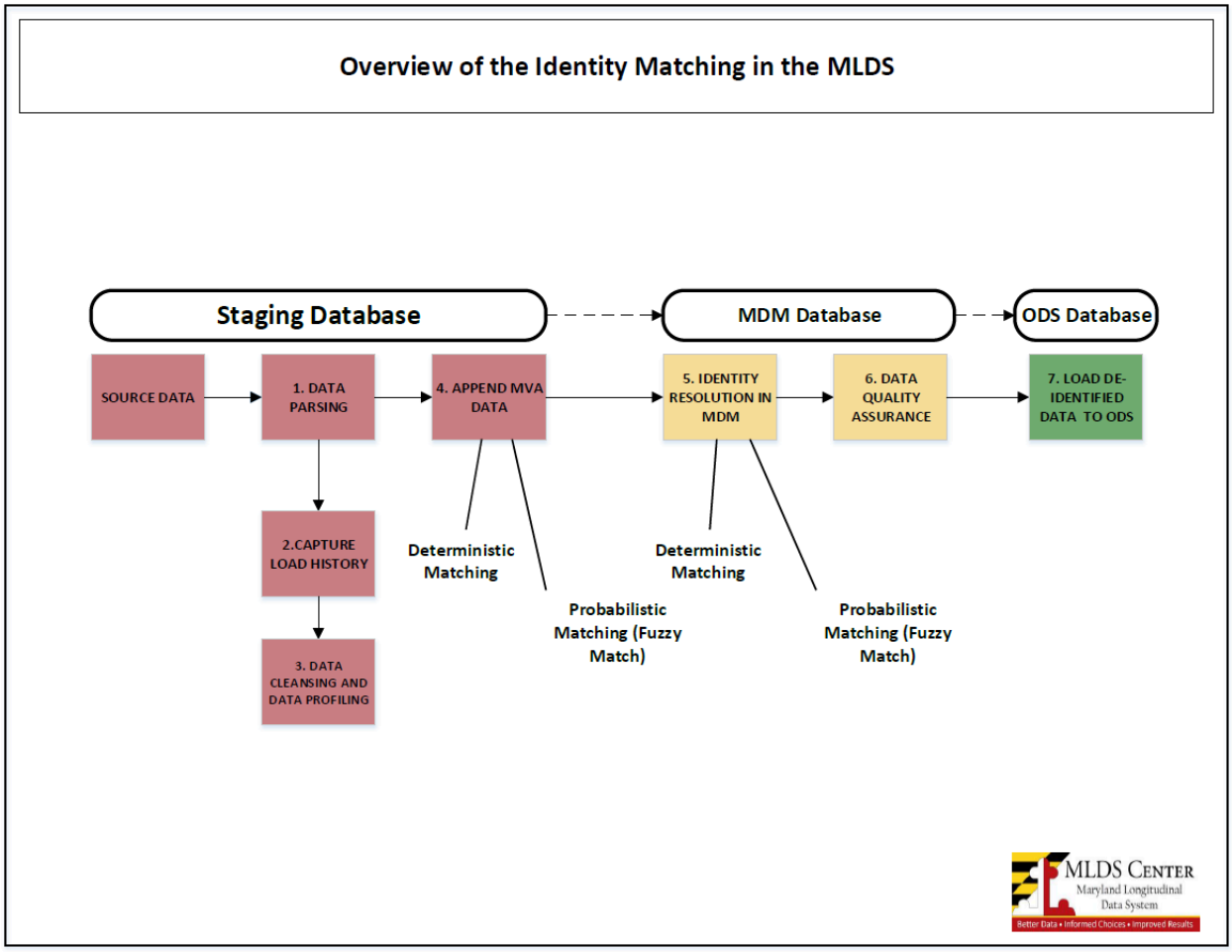
III. The Step-by-Step Process for Loading Source Data in the MDM

This section documents key steps in the data loading process which is a critical component of conducting identity resolution. This does not include details on the ODS loading process.

Data Process Flow

Figure 1 represents the components of the MLDS and the flow of data through the System starting with an agency collection source file and ending with data ready to be loaded to the ODS. Each of the components are described in detail further in the document.

Figure 1: Overview of the Identity Matching in the MLDS



Step 1: Data Parsing

The first step is to parse the file and compare the format of the file received against the expected format (e.g. .csv, .txt). The expected format is provided with an agency collection source file specification document, which is to be included with each file uploaded by the agency partner. The file format specifies not just the file type, but the order (for .csv) or length (for flat files) of each field in the file. If the format of the provided file does not match the expected format, the Center reports the issue to the data partner agency liaison and awaits a corrected file.

Step 2: Capture Load History

The Center then loads the file to a temporary staging area in the MDM and assigns a source sequence number, which is a unique identifier for each row. If needed, this source sequence number aids in referring back to the original source record. The Center also assigns a Load ID and Load Date. The Load ID and Load Date assist the Center in keeping an audit trail of the file load. Once all this information is assigned, the Center documents preliminary counts of records and unique individuals within the MDM database. As part of this process, data quality steps are applied such as confirming that the number of rows of data in the MDM match the source file and expected values are present in the source file.

Step 3: Data Cleansing and Data Profiling

The next step is to load the file to the MDM in temporary tables (also known as the staging area) and cleanse the data in preparation for identity resolution. This entails analyzing and reviewing the data, to identify missing values or types of errors that could impact the structure of the dataset. The data cleansing process predominantly focuses on names, but also examines dates (ex. Date of Birth) for differences in format.

Conducting name cleansing entails isolating any generational suffixes (Jr, Sr, II, III, etc.) that exist in the first, middle, or last name. This information is then separated into a Generational Suffix column. The Center uses custom built functions in Oracle to do the name cleansing. This enables the Center to make edits to the process based on new collections that may have different formats or data fields, if needed. The code is only looking for a valid format so if it is not in the valid format (number or symbol), the field becomes null. (ex. 00000 as name, would become null). Dates are checked for the proper format (DD-MMM-YYYY). For example, if a date reads April 1, 2020 the code will reformat the date to 01-Apr-2020; if only April 2020 (month and year) is provided it will be flagged and not converted into an absolute date. Date values are used in the matching process.

Step 4: Using MVA Data

Prior to resolving identities in MDM, the MLDS Center uses MVA data because it includes a Social Security Number (SSN) that is verified with the Social Security Administration and the information can be used in future matching attempts with new agency collection source data. The next step is to compare data in the staging area to MVA data.

The agency collection source Personally Identifiable Information (PII) data includes the following:

- First Name
- Last Name
- Middle Name (Complete or Initial)
- Generational Suffix
- Date of Birth (Complete or Year/Month)
- Social Security Number

This part of the process depends on the PII data received by the Center in the agency collection source file because not all files have complete PII data. If the agency collection source file does not contain one of the PII data elements above it is left blank in the table.

Next, the Center uses MVA attribute data by matching agency collection source Social Security Number, Names, and Date of Birth.

The MVA columns include:

- First Name

- Last Name
- Middle Name (Complete or Initial)
- Generational Suffix
- Date of Birth (Complete)
- Social Security Number
- Gender

Deterministic Matching for Using MVA Data

On the next page is the order in which Center data analysts conduct deterministic matching with MVA data. Refer to Table 1 to see the order in which the matching algorithm proceeds. The algorithm begins with the strictest matching conditions; i.e. Order 1 must match exactly on ALL 6 fields: DOB, Last Name, First Name, Middle Name (full name or initial), Generational Suffix (Gensx), and SSN. If a match cannot be made using the conditions from Order 1, then the analyst proceeds to the conditions of Order 2, and so on. When a match is found, the analyst records the Order # in the table in the MDM used to satisfy the match conditions. Each record gets assigned an Order ID so the Center can capture and report on the count found per match order.

Once a match is found, all MVA PII data gets appended to the source record. No source data is overwritten. Also, once a match is found, the search for a match ends and does not proceed further. If no match is found, the associated MVA columns will be NULL.

Currently, the exception in the matching order is that UI Wage data is matched to the MVA data on SSN only (see order #7 in Table 1). If a match on SSN is found, the additional MVA data (DOB, name, etc.) is used. For these matches, the next step is conducting the deterministic matching process within the MDM (see page 9).

Table 1: Deterministic Matching Order for Using MVA Data

X = Exact

F: First (ex. F3 = First 3 letters)

Gensx: Generational Suffix

Order	DOB	Last	First	Middle	Gensx	SSN	Requirement
1	X	X	X	X (full or initial)	X	X	EXACT on 6
2	X	X	X	X (full)		X	EXACT on 5
3	X	X	X	X (first initial)		X	EXACT on 5
4	X	X	X			X	EXACT on 4
5	X	F3	F3			X	EXACT on 4
6	X					X	EXACT on 2
7						X	EXACT on 1
8	X	X	X	X	X		EXACT on 5
9	X	X	X	X (full)			EXACT on 4
10	X	X	X	X (first initial)			EXACT on 4
11	X	X	X				EXACT on 3; extra review process

Probabilistic Matching for Using MVA Data

Data that do not match deterministically (based on exact matches of fields) are matched probabilistically (based on close matches of fields). The Center utilizes the Levenshtein Method and the Oracle utl_match.edit_distance utility to perform probabilistic matching and determine the distance between strings.

Within each step in the table on the next page, based on the results, we assign a rank and the higher the rank the higher the confidence level in the match. Each probabilistic match goes through manual review. The highest order rank is selected and established as a match.

Table 2: Probabilistic Matching Order for Using MVA Data

X: Exact
ED: Edit distance
F: First (ex. F3 = First 3 letters)
Gensx: Generational Suffix

Order	DOB	Last	First	Middle	Gensx	Gender	SSN	Req
1a	X	ED=1	ED=1			X		
1b	X	ED=1	ED=2			X		
1c	X	ED=2	ED=1			X		
2	X	F3 & ED	ED			X		
3	X	ED	F3 & ED			X		
Date of Birth Errors:								
4	ED	F3 & ED	ED			X		
5	ED	ED	F3 & ED			X		
6	ED					X	X	

Step 5: Identity Resolution Process in MDM Database

Using the source file and matching MVA data (if available), the Center matches against the MDM database (PARTY Model) to determine whether the identity already exists in the MDM. If a match is found, the existing PARTY ID associated with an identity in the MDM database will be assigned and appended to the identity in the source file.

The Center assigns a PARTY ID by forcing minimum and maximum possible values of the PARTY ID. If the minimum and maximum value are the same, the Center uses the existing PARTY ID. If the minimum and maximum values are not the same, the Center confirms which value should be merged or unmerged. This process is explained in further detail below.

The MLDS Center uses records from specific source files to establish identities in the MDM and create PARTY IDs. Source files are currently received from the following data sharing partners:

- Maryland State Department of Education (MSDE);
- Maryland Higher Education Commission (MHEC);
- Maryland Department of Labor (Labor);
- Maryland Department of Juvenile Services (DJS);
- Maryland Department of Human Services (DHS);
- Maryland Department of Health (MDH); and

- Other supplemental sources, such as national certification providers.

Deterministic Matching for MDM

Below is the order in which the Center conducts deterministic matching between the source file and any MVA data with the MDM (PARTY Model). In other words, if no MVA data was found for an identity the source file information is still matched against the MDM. Refer to the chart below, to see the order in which the matching algorithm proceeds. The algorithm begins with the strictest matching conditions; i.e. Order 1 must match exactly on ALL 7 fields: DOB, Last Name, First Name, Middle Name (full name or initial), Generational Suffix (Gensx), SSN, and Source Unique ID. If we are unable to find a match using the conditions from Order 1, then we proceed to the conditions of Order 2, and so on. When a match is found, we record the Order # used to satisfy the match conditions. Each record gets assigned an Order ID so the Center can capture and report on the count found per match order.

Table 3: Deterministic Matching Order for Matching Identities

X = Exact

F: First (ex. F3 = First 3 letters)

Gensx: Generational Suffix

Order	DOB	Last	First	Middle	Gensx	SSN	Source Unique ID	Requirement
1	X	X	X	X (full or initial)	X	X	X	EXACT on 7
2	X	X	X	X (full)		X	X	EXACT on 6
3	X	X	X	X (first initial)		X	X	EXACT on 6
4	X	X	X			X	X	EXACT on 5
5	X	F3	F3			X	X	EXACT on 5
6	X					X	X	EXACT on 3
7						X	X	EXACT on 2
8							X	EXACT on 1
9	X	X	X	X	X			EXACT on 5
10	X	X	X	X (full)				EXACT on 4
11	X	X	X	X (first initial)				EXACT on 4

12	X	X	X					EXACT on 3; extra review process

Probabilistic Matching for MDM

As stated above, the majority of matches are found using deterministic matching, but for the remaining unmatched data the Center uses probabilistic matching (also known as fuzzy matching). The Center utilizes the Levenshtein Method and the Oracle util_match.edit_distance utility to perform probabilistic matching and determine the distance between strings. Note: Unique Source IDs are not part of probabilistic matching because all Unique Source IDs would have been matched in the deterministic matching process.

Within each step in the table below, based on the results, we assign a rank and the higher the rank the higher the confidence level in the match. Each individual record goes through manual review. The highest order rank is selected and established as a match. If more than one existing identity matches at the same rank order, then the subsequent matching order is followed and identities may be merged or unmerged.

Table 4: Probabilistic Matching Order for Matching Identities

X: Exact
ED: Edit distance
F: First (ex. F3 = First 3 letters)
Gensx: Generational Suffix

Order	DOB	Last	First	Middle	Gensx	Gender	SSN	Req
1a	X	ED=1	ED=1			X		
1b	X	ED=1	ED=2			X		
1c	X	ED=2	ED=1			X		
2	X	F3 & ED	ED			X		
3	X	ED	F3 & ED			X		
DOB Errors:								
4	ED	F3 & ED	ED			X		
5	ED	ED	F3 & ED			X		
6	ED					X	X	

Resolving Identity Conflicts

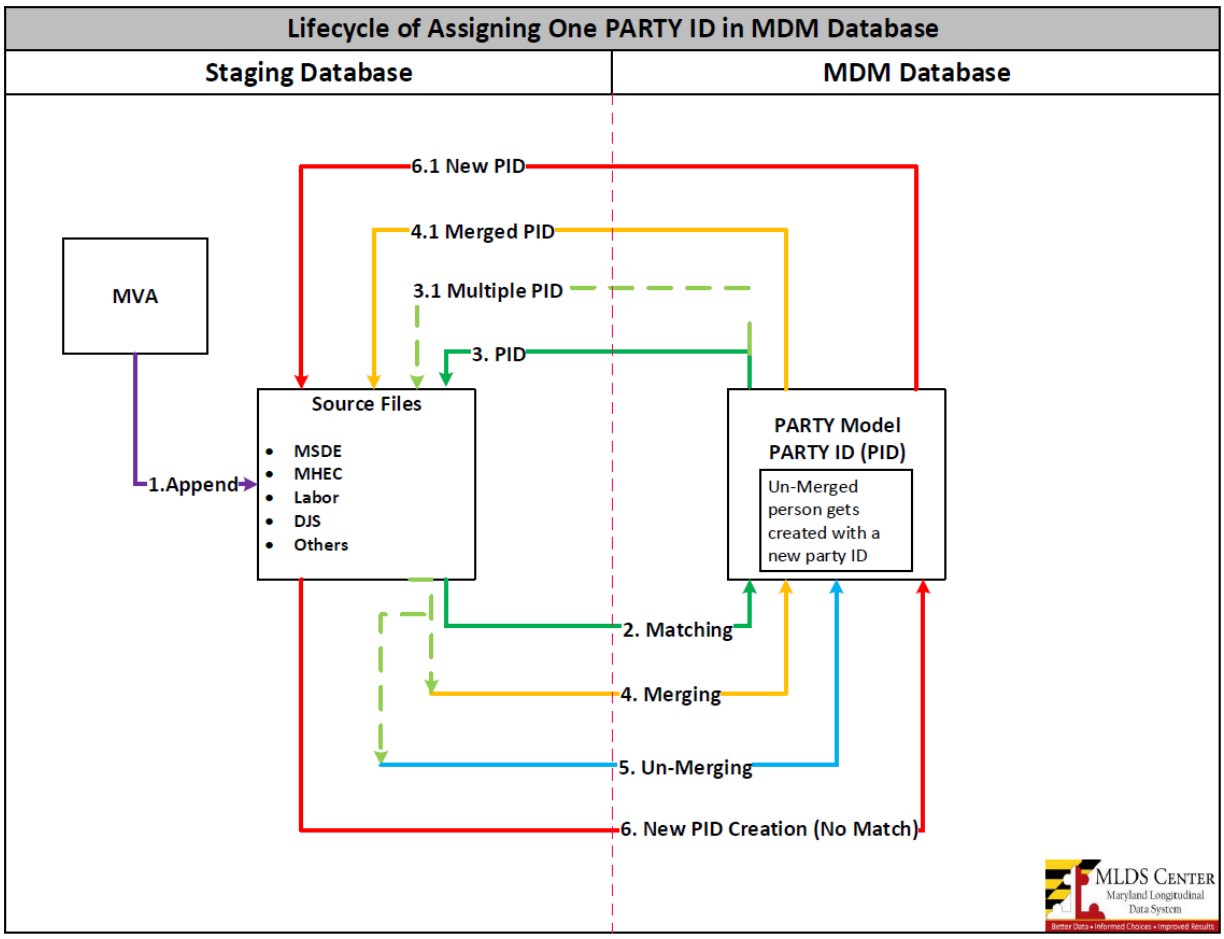
When probabilistic matches are found, if any differences exist, the assignment of an identity is diligently performed by checking other sources. Based on the manual review results of the conflict, the Center will determine whether an error has occurred and, if so, the nature of the error. One type of error occurs when two PARTY IDs exist when in fact there is only one identity. This may occur because demographic information may not have been available at the creation of a particular PARTY ID. Another type of error occurs when two distinct identities are associated with one PARTY ID. This may occur because of an error in source file data. To resolve the error, the Center will either **merge** the identity or **unmerge** the identity.

- **Merge:** If there are two PARTY IDs for the same identity in the MDM, the Center merges the PARTY IDs into one PARTY ID. The minimum PARTY ID (earliest created PARTY ID) is retained as that PARTY ID.
- **Unmerge:** Separating two identities from one PARTY ID.

Creating New Identities

The Center confirms an identity never existed in the MDM using a thorough manual review process as described above. In the cases where no identity can be matched in the MDM the Center creates a new PARTY ID for that identity in the MDM.

The following diagram illustrates how PARTY IDs are assigned in MDM:



Step 6: Data Quality Assurance

Once identity resolution has been performed, several additional data quality steps are in place to mitigate the source reporting errors, outliers, or any specific code exceptions in the source fact data.

As part of quality assurance, all the fields in the source files are compared against the approved file formats to ensure the values and formats. If there are any approved score value ranges, those are again checked at this stage. The Source to Target Mapping documents are created to document where every field from source files are loaded in the Operational Data Store.

The Center goes column by column to identify any potential issues, such as:

- Making sure that all values are permitted values.
- Identifying reporting errors as described in the source manual.
- Noting inconsistencies between source data and business rules / logic.

If an issue or question arises, the MLDS Center uses an internal 'Data Inquiry Form' process to review and create relevant documentation. These issues are also discussed with the data sharing partners, internal Data Team, and the Data Governance Advisory Board.

Step 7: Load De-Identified Data to ODS

Once the data quality step is performed, the de-identified data, with PARTY ID along with the facts are loaded into the MLDS Operational Data Store.

IV. Glossary/Definitions

- **Data Cleansing** - Reviewing and focusing on identifying missing values or types of errors that could impact the structure of the dataset.
- **Data Parsing** - Comparing the received file format against the expected format
- **Data Profiling** - The process of examining the data available from an existing information source and collecting statistics or informative summaries about that data.
- **De-identified Data** - Data that has had all PII data elements removed. De-identified student data could not be used to figure out the identity of the student that the data describes.
- **Deterministic Matching** - Computerized comparison where all criteria need to match exactly.
- **DJS** - Department of Juvenile Services
- **DHS** - Department of Human Services
- **Identity** - A unique person record in the MDM
- **Identity Creation** - Establishment of a new PARTY ID in the MDM
- **Identity Resolution** - Identity Resolution; the process of connecting disparate data sources to determine whether identities of two or more records are the same
- **Labor** - Department of Labor
- **Edit Distance/Levenshtein distance matching method** - A measure of Similarity between two strings, s1 and s2. The distance is the number of insertions, deletions or substitutions required to transform s1 to s2.
- **Load ID** - The ID number given when a data source is loaded to the MDM.
- **Load Date** - The date a data source is loaded to the MDM.
- **MDM** - Master Data Management database. This restricted access database system contains unit-level PII data and identities created from source files.
- **Merging** - When there are two PARTY IDs for the same identity in the MDM, the PARTY IDs are combined into one.
- **MHEC** - Maryland Higher Education Commission
- **MSDE** - Maryland State Department of Education
- **MVA** - Motor Vehicle Administration. MLDSC receives MVA data and uses it for matching and appending to source records as needed.
- **ODS** - Operational Data Store. This database system contains deidentified records with all data collection information.
- **Partner Agencies** - MSDE, MHEC, Labor, DHS, DJS, MVA.
- **PARTY ID** - The unique identifier for an identity generated in the MDM.
    - Minimum PARTY ID: created first

    - Maximum PARTY ID: created after
- **Person ID** - A de-identified unique ID assigned to a person in the ODS
- **PII** - Personally Identifiable Information (PII Data): This term refers to data that could be used (by itself, or in combination with other sources of data) to uniquely identify an individual.
- **Probabilistic Matching** - A statistical approach to measure the probability that two records are for the same individual.
- **Staging Area** - The temporary storage area between a source file and the MDM
- **Source Data** - The administrative data collected by state agencies and submitted to the MLDS Center by partner agencies. (Ex. MSDE provides Attendance data for students in K-12, MHEC provides Enrollment data for students attending Maryland's community colleges, four-year public institutions and state-aided independent institutions, Labor provides wage information for Maryland employees. Source data also includes other data sources, such as national certification providers.
- **Source Unique ID** - A unique person identifier generated by the data provider.
- **Unmerging** - Separating two identities from one PARTY ID.
- **Verified Source** - The data provided has been checked against a database for accuracies and inconsistencies