About Gaine Solutions

Gaine is a specialist in Enterprise Data Management (EDM) with a particular focus on Master Data Management (MDM) and Master Data Governance (MDG).

Since 2007, Gaine has worked closely with some of the world's largest organizations including healthcare providers, semi-conductor manufacturers, Property and Casualty insurers and Fortune 500 biotechnology companies. We help them integrate, improve and govern their Master Data assets.

Gaine's MDX (Master Data eXchange) software platform offers our customers the power to identify critical information and create a single view of their data. Gaine has integrated master data from more than 2,000 operational systems covering a wide variety of application technologies.

Master Data Management

There are many types of data. Two major types are **transactional data** and **master data**. Though these types aren't mutually exclusive, it helps to think of one in contrast to the other.

Master Data refers to the largely-static sets of data that are critical to a business. For example, a grocery store will have lists of **products**, **customers**, and **suppliers**. There are a finite number of entries in these sets, and they generally only change slowly over time.

Transactional Data, in contrast, represent facts about the world that relate these master lists together, typically at a specific point in time. For example, "Customer **John Smith** bought 6 **Washington apples** from Vons at 2:00 on Tuesday" is a transaction. Transactional lists can grow indefinitely.

Everyone deals with the problem of maintaining master data. To illustrate, consider the list of contacts on your smart phone

- **Keeping lists accurate** as data changes over time. For example, your friend John changed his cell number when he moved, but your phone still has the old number. This problem is the problem of **slowly changing dimensions**.
- **Cleansing** data. For example, you entered some numbers as (###) ###-#### and others as ###.###.####, making it hard to identify duplicates.
- **Deduplicating** data, particularly from multiple sources. For example, your phone integrates your Google contacts with the contacts on your SIM card and those stored on the Verizon servers, so now your friend Paula appears three times.

These problems are compounded when working with an enterprise's multi-million row databases. **Master Data Management** refers to this process of maintaining and centralizing an enterprise's master data lists.

General Overview of an MDM Process

- **ETL** (Extract, Transform, and Load) – Data is **extracted** from one or more source systems, **transformed** into a new data model that supports master data management, and **loaded** into an MDM database

- **Cleansing** – Data is cleansed and standardized. The business must establish their accepted standards and enforce these guidelines. For example, it's common to apply USPS standards to addresses, changing things like '1ST Street' into '1ST ST'. This improves the quality of the data and makes matching easier.
- **Matching –** Fuzzy matching is run to identify duplicates. This allows records that are similar, but not exactly the same, to combine together.
- **Master Key Assignment** – Each contributing record is assigned a 'master key' based on the results of the matching effort.
- **Survivorship –** After records are matched together, the question remains of which record has the *best* data. Survivorship is the process of applying a set of rules to identify the best data from all matched records, in order to build a new 'master record.' This is different from simply picking one of the records as the 'winner', because the master record may have elements from all contributors.
- **ETL –** Data is extracted from the MDM database, transformed into the destination format, and loaded into the target system.

The Scenario

You are working with an insurance company that gets thousands of medical claims on a daily basis. A claim is a bit of transactional data, saying which **member** (patient) met with which **provider** (doctor), what services were performed, and how much was billed.

To reduce the risk of fraud, this insurance company pulls reports every quarter on which providers and which members submitted the most claims. This can help shed light on suspicious activity, like an unusually high number of claims from a single doctor.

However, the insurance company has recently gone through several acquisitions. Each acquisition brought with it a separate database of providers, in a different structure. These databases are currently living side-by-side, and must all be combined in order to pull an accurate report.

When the databases are combined, it creates a lot of duplication and poor data quality. This is making it difficult to obtain a 'single view' of the provider. If the sketchy Dr. Ozwald submits claims under the names Dr. Ozwald, Dr. Oz, and Dr. Ozzy, then the insurance company might not catch his nefarious activities.

They would like to do a cleanup effort on their provider list that includes **de-duplication**, **cleansing**, and finding a '**best version of the truth**.'

The Task

Your task is to create a basic MDM engine to help them achieve this goal.

Project Requirements:

- Need to support a **one-time data load** of data. We will provide the data.
- Must assign an '**MDM ID**' and be able to trace each source record to its master.
- Must have **complete auditing** on why specific consolidations were made.

- Must build a master record that contains the **“best” attributes** from all contributing records. These might not all come from one record.
    - Judgments about what make up the “best” data are left to your discretion, but must be auditable
- Ability to **have client-configurable match guidelines**
- Ability to **extract data in a flat format** TBA (so that we can test it).
- Stretch goals:
    - Some kind of useful **visualization** on how a master record is composed.
    - Ability to search and browse the master records via a **GUI**.

Data Specification

Input Data

The data we will provide is essentially a list of **medical providers.** A provider could include both individual practitioners (like Dr. Jones) and organizations (like ‘Allergy Partners of the Central Coast’ or ‘Kaiser’). All information except for the provider’s type and name may not be present. Provider information includes:

- **Type** – We will provide the type of provider for you: "Individual" or "Organization". This is something that is not always obvious because of what an "organization" actually is. It can be a very formal organization like the American Public Health Association, a looser collection of doctors that practice together, or even something like a sole proprietor who owns their own practice. In the latter case, it is not unusual to have a practice that is the same as the doctor name. Therefore, there may be an entry for both "Dr. John Smith" the doctor and "Dr. John Smith" the practice. These are two distinct entities that **SHOULD NOT** be merged together.
- **Name** – This is a single name field, which may include first, middle, last, prefixes, suffixes, and credentials.
- **Gender** –The gender of the provider as 'M' or 'F'.
- **Date of Birth** – The date of birth of the provider. Different date formats are possible.
- **Is Sole Proprietor** – Whether or not this provider is a sole proprietor. A sole proprietor is a doctor who owns their own practice. Possible values are 'Y' (yes), 'N' (no), or 'X' (not answered by provider).
- **Mailing Address** – The address that the provider receives mail at. Note that it is not uncommon for providers to receive mail at PO boxes. See below for what an address may contain.
- **Practice Address** – The address that the provider actually provides services. See below for what an address may contain.
- **Phone** – The primary phone number for the provider. Phone numbers can be in various different formats.
- **Specialty** – A provider's specialty is the medical field that provider specializes in. This is often also called a taxonomy. A provider can be a specialist in several different (often related) fields. For example, it would not be unusual for a pharmacist to have a specialty of both 3336C0003X (Suppliers/Pharmacy, Community/Retail Pharmacy) and 333600000X (Suppliers/Pharmacy). Although a provider can have many specialties, each source record only has two specialties provided: Primary and Secondary.

The following addressing information may be present:

- **Street** –The street address. Typically the first line in an address.
- **Unit** –The unit/suite/building/apartment/lot/etc number.
- **City** – The city.
- **Region** – The region of this address. If in the US, this will be the state or province. Don't forget that there are more than just states in the United States.
- **Post Code** – May be the five or nine digit post code.
- **County** –Name of the county or regional district.
- **Country** – Name of the country. This will be the full name, not an ISO code.

Output Data

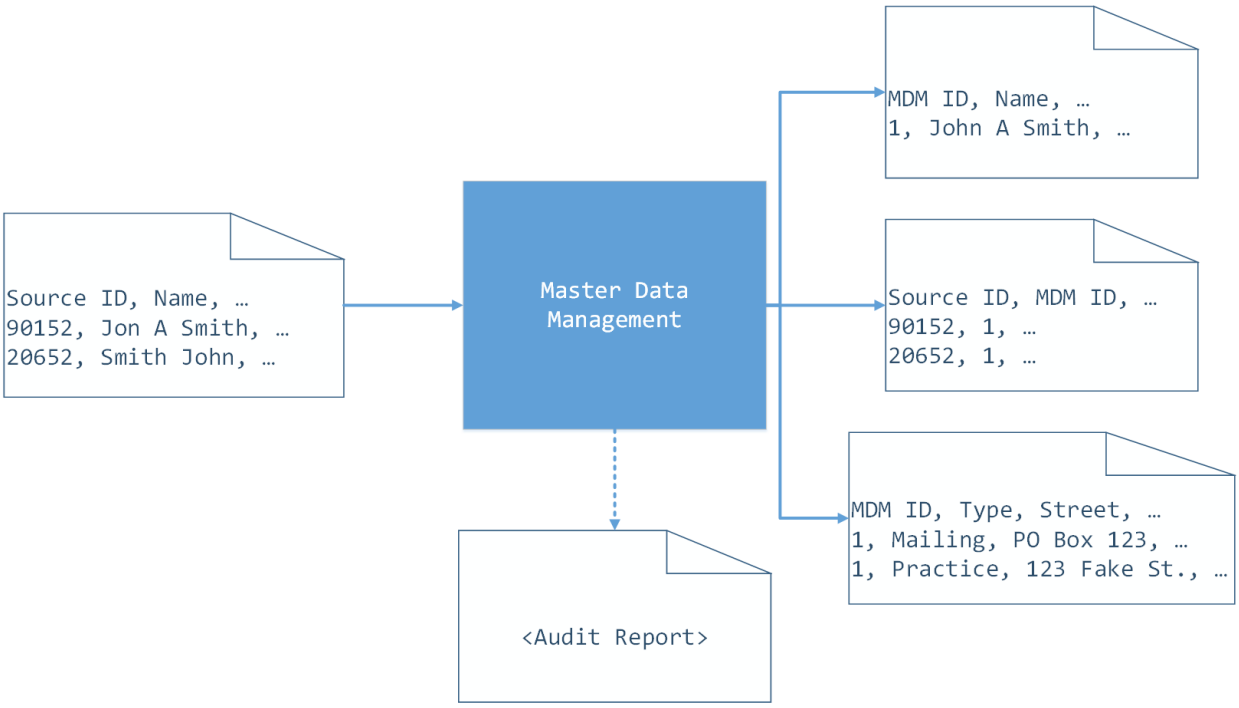
For the data returned to us, we will use what we will call the “Master Crosswalk” scheme. The exact extract specification will be provided once the data is released.

Two files will be used to capture mastered provider information:

- **Master File** – Contains the mastered (best collection of data) records for each provider along with the MDM id. This file will contain the provider's type, name, gender, date of birth, sole proprietor status, phone, and specialties.
- **Crosswalk File** – This file is used only to link source records to mastered records. It will just contain the source id and MDM id.

Addresses will be in a separate file that has the MDM id, type of address (mailing or practice), and addressing information. Duplicates (non-mastered addresses) are acceptable for the address file.

In addition, you will be asked for an audit report that details the decisions made for each consolidation and survivorship. The format of this file will be on a per-team basis.



Release of Data

A subset of the data along with the full output specification will be released a few weeks into the quarter. A few weeks later, the full dataset will be released.

Result Scoring

We will run your results through a scoring system to help evaluate your system.