Master Data with DataStudio

How to Automate Data Reconciliation for Large, Multi-Variate Data Sets

Background

All data that a business needs is seldom from a single source. Different data sources might capture different facets or attributes pertaining to a particular entity as shown in Figure 1. Some of the common attributes between sources might also present conflicting information about the entity. Consider an event related to catastrophic equipment failure for an industrial company. Information about this failure might be recorded in a variety of databases. For example, the maintenance database might record information such as the date of failure, root cause and preventive actions taken to address the failure, while a different safety events database might record the injuries, personnel involved, compensation paid and other safety-related meta data.

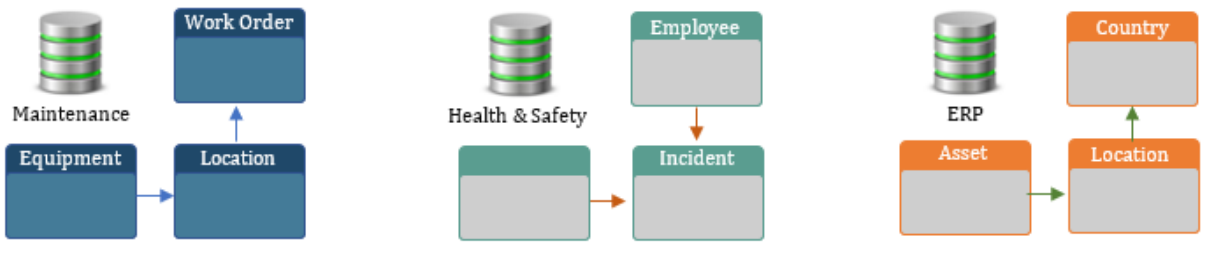

Figure 1: Disparate Data Sources

Frequently, disparate data sources do not have unique identifiers that map events or entities between them. That is, the failure event might have two different identifiers in the maintenance and safety databases. As a result, it becomes too difficult to answer questions like, "what is the true cost of the equipment failure that caused the injury?"

Most companies do not have the resources and budget to generate a common schema for enterprise entities which could cater to multiple departments including operations, finance, marketing, accounts, etc. IT departments may make an attempt to do so, oftentimes unsuccessfully. This is because implementing an enterprise Master Data Management (MDM) systems can be very expensive and involves significant manual work. The complexity of the project goes up exponentially if IT has inherited several disparate systems through past mergers and acquisitions. To this end, a common need of industrial data users is to integrate data from multiple databases, where data pertaining to the same entities is collected from various sources, reconcile conflicting information and aggregate complementary information.

Traditional Approach

The enterprise data integration and reconciliation task typically involves multiple steps as shown below:

Data Deduplication	Mapping	Reconciliation	Aggregation:
Identifying the set of unique entities or events across the data sources	Associating each unique event with a global identifier and mapping each entity in each source to the global identifier	Resolving conflicts regarding the attributes of the data using prioritization rules	Collecting the multiple attributes presented in the various sources to form a global set of attributes for each unique event.

In the IT world, this process is often called data mastering, and solved using software such as Informatica MDM. The master data collects data from various sources and identifies identical or similar data and creates a database which becomes the single source of truth.

This whole process involves significant manual work and is usually performed by a data expert using software. This manual process is also custom based, i.e. it greatly depends on the data sources and hence, needs to be altered as the sources change. Also, when the data includes geospatial components and text, which is not in the same format in all sources, the process becomes even more difficult.

The DeepIQ Approach

DeepIQ has developed patent pending, AI based automated data reconciliation software that uses the power of distributed computation and machine learning to automate the process of master data discovery.

DeepIQ's software uses a combination of statistical and machine learning algorithms to achieve a high-quality data mastering process. First, a series of statistical algorithms process all the sources to identify all versions of an entity or event across sources. Next, the machine learning algorithms refine the output of statistical algorithm to prune out poor quality results and generate high fidelity mapping of sources to globally unique entities. A final prioritization or fusion approach amalgamates all sources to create golden records of each unique entity.

These sequences of algorithms can be very computationally intensive. Consider for example, creating a material master for a combination of two sources, each with a million records. A naïve implementation to identify the unique entities in the union of the two sources would require a trillion computations, where each entity in a source is compared to every entity in another source.

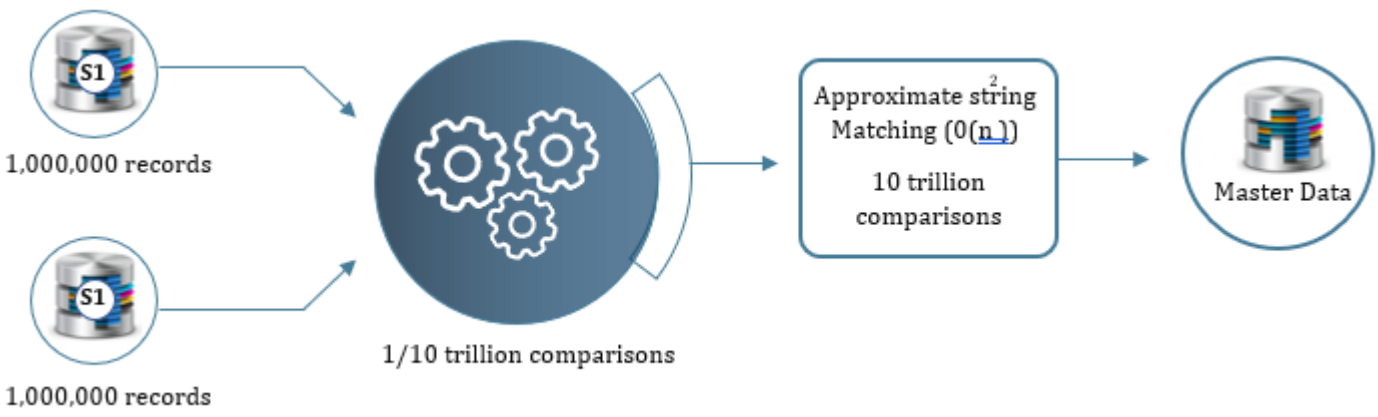

Figure 2: Comparisons of millions of records

DeepIQ uses two strategies to handle this computationally intensive task. First, it runs natively on a distributed computational platform that supports Adobe Spark. Based on these techniques, DeepIQ's software is able to run the end-to-end master data workflows in just a few minutes even dealing with sources with multiple millions of entities.

DeepIQ's DataStudio provides an easy to use, drag and drop interface for the relational ingestion and data preparation like fixing time zones and filtering. DataStudio enables the user to create simple workflows for Master Data.

`The Spark code is generated from a combination of the user generated workflows and the DeepIQ libraries. This code is passed to Spark execution.

Success Story

An E&P company has grown through mergers and acquisitions. As a result, it has multiple data sources from different organizations within the company containing information on all of the materials that it procures, stores and uses. The original material master was no longer complete, and all downstream applications suffered from duplicate material records, conflicting material configuration data and incorrect stock estimates.

The company decided to drive optimal processes in purchasing and inventory management to minimize stock outages or unnecessary purchases. A traditional master data management solution would have taken more than a year to complete with multiple people and several technology stacks. The company conducted a pilot with a leading vendor that used a machine learning process to automate this task. The software license costs alone were in excess of $1 million per year and yielded mixed results at best.

At this point, the company started an engagement with DeepIQ to automate this process and achieve immediate results. DeepIQ installed its DataStudio software and configured it to run on customer's existing infrastructure. Within four weeks, using DataStudio, our client configured the workflow to:

- Ingest the material master data from three different sources
- Perform cleansing operations to ensure data quality
- Create golden records for materials across all customer sources
- Create references to the master data records in the multiple datasets

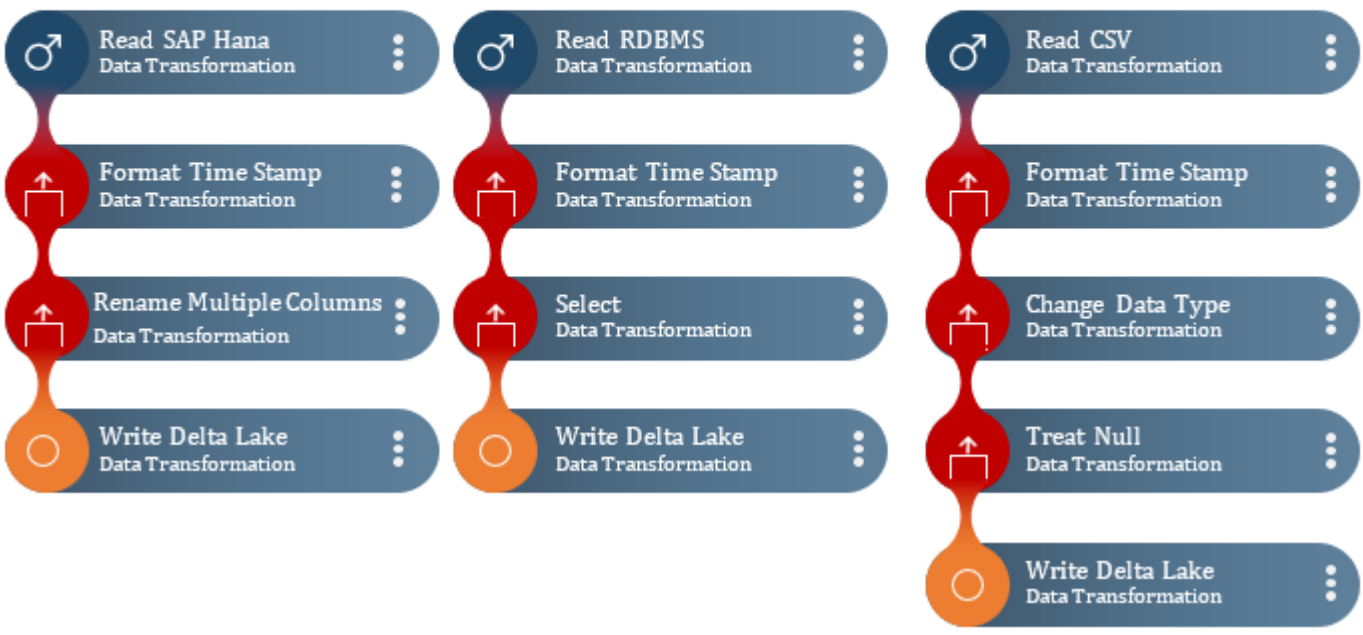

Figure 3: Sample Master Data Workflow

The solution was designed so that as new records were created, updated or deleted in the source systems, the master data was updated. Processing all three data sources, including deduplication of the data within each source, took only 12 minutes to complete on a medium sized cluster.

To ensure data quality, any records deemed to be between 75% to 90% match were sent to a human data steward for verification. Random matches were also sent to the data steward. The decisions made by the data steward were used to further train the master data model enabling the matching algorithm to improve over time.

The following diagram shows a sample of a typical match and the results.

Model	C-018	C-018X	C-118
Description	133/8 stop collar	13.375 St. collar	13.3/8 shaft collar
Dimension	3/4X3/8X3/8	.75x.375x.375	3/4X3/4X3/4
Purchase Date	NULL	5/1/2010	6/15/2015
Material Group	Casing Centralizers	Casing Centralizers	Casing
Manufacturer	ABC Inc.	ABC Incorporated	ABC Inc.
Match Score		90%	40%

Figure 4: Sample Match Results

Conclusions

DeepIQ's patent pending Master Data capability captures the complex interactions between different data entities even when the training data is sparse. This unique capability can be a gamechanger for your enterprise data, allowing users to more robustly analyse their business using the full scope of their dataset.

https://www.informatica.com/products/master-data-management.html