Gain Advanced 360 Degree Views of Your Data

A Technical Guide to Novetta Entity Analytics
NOVETTA
ENTITY ANALYTICS
Gain Advanced 360 Degree Views of Your Data

CONTENTS
1 INTRODUCTION
1 MAPPING AND PROFILING DATA
4 MEASURING AND TUNING DATA
4 ENTITY GROUPING AND RESOLUTION
5 ENTITY LINKING AND RELATIONSHIP ANALYSIS
6 PUBLISHING RESOLVED ENTITIES AND RELATIONSHIP RESULTS
7 ACCESSING ADDITIONAL CAPABILITIES
7 NOVETTA ENTITY ANALYTICS: DELIVERING POWERFUL ENTITY AND RELATIONSHIP INSIGHTS
INTRODUCTION

Novetta Entity Analytics combines data from multiple sources and produces advanced 360-degree views of real-world entities and their relationships to other records and entities. The high-fidelity entity information Novetta Entity Analytics produces is used by other applications and workflows to enable users to perform new types of analyses and gain insights about people, organizations, events, and other entities that aren’t possible with other technologies. Novetta Entity Analytics is used by commercial and government organizations that want to perform advanced customer 360 analytics; detect risk, fraud and abuse; and assess security threats.

This white paper provides technical readers with a walkthrough of the processes Novetta Entity Analytics uses to combine data from multiple sources and rapidly resolve and publish information about real-world entities contained within that data. The workflow processes and other information covered are mapping and profiling, measuring and tuning, and grouping and resolving data; detecting and linking relationships among entities; publishing entity and relationship information; and other technical details.

MAPPING AND PROFILING DATA

The first stage in creating resolved entities within Novetta Entity Analytics is to map data sources to a common data model based on the attributes of all of the data being combined. This stage also includes profiling individual and combined data sources to make it easier for users to accurately resolve data. The steps involved in the mapping and profiling stage are as follows:

Create Workspace, and Select Workflow and Entity Type

Users first create a new workspace folder to store their work and select the workflow and entity type appropriate for their specific project. Novetta Entity Analytics currently supports grouping and linking workflows and has built-in rules for person, organization and location entities, (see Figure 1). Additional entity workflows and entity strategy rules can be added from the application’s developer interface.

Figure 1. Create Project Workflow
Users choose either the grouping or linking workflow; select a person, location or organization entity type; and enter a workflow name for their specific project within the create workflow dialog box.
Once workflow and entity types are selected, the application opens a window that contains all of the applicable strategy rules for matching and resolving or linking entities for the specific project. In Figure 2, the built-in strategy rules for grouping person entities are shown in the left-hand column.

**Add and Profile Data Sources**

Next users select and add the data sources they want to use for their project. Novetta Entity Analytics works with any data source registered in HCatalog and can scale to handle a large number of sources and records. The software is currently in use on projects that involve hundreds of data sources and billions of records.

Once a data source is loaded, users can automatically profile the source information and view details about the top values and patterns within individual fields, see a profile of name patterns in Figure 3. This information helps data scientists and other users better understand the contents of each data source, so they can determine how best to cleanse and map the data.

During this step, Novetta Entity Analytics also runs data profiling and mapping jobs in the background using Spark or MapReduce, depending on the organization’s environment, and stores the patterns and values it produces in Hive for the application’s future use.

**Cleanse Data Sources**

Novetta Entity Analytics includes many built-in cleansing functions designed to improve data consistency and accuracy when applied to source data. Novetta’s team developed the cleansing functions based on their work with U.S. government agencies preparing data sources for entity resolution.

Users select the functions they want to apply to source data, based on what they know about the characteristics of the data they are using for their project (see Figure 4). If needed, users can also add their own custom cleansing functions from the application’s developer interface.
Map Data Sources, Calculate Statistics

Once cleansing functions have been selected, Novetta Entity Analytics applies them and automatically maps data sources to the workflow’s entity data model. Based on the attributes contained within source data sets, the software identifies and pre-selects the best strategies to use when grouping entities. In Figure 5, the strategy rules are in the left hand column and the best strategies are dark green. Users can also add custom data models or strategy rules from the application’s developer interface.

Once Novetta Entity Analytics maps the data sources, it creates a unified table containing all of the data. The software also performs follow-on jobs in the background to calculate statistics about the combined sources, e.g. how many phone numbers does a specific person have, or how many different names are associated with a single phone number. These statistics are then added to the unified table, as new attribute fields, and are used to calculate uniqueness values during the measuring and tuning, and grouping stages.

Profile Unified Data

Novetta Entity Analytics can also quickly profile data fields across all data sources to provide users with additional insights about the values and patterns within the combined data sources. This information helps users optimize data during the measuring and tuning stage. In Figure 6, a profile of phone number values is provided.

Standardize Addresses

In this stage, if required, users can also easily apply standardization procedures to address data. Novetta Entity Analytics uses a library provided by the United States Postal Service to check addresses for validity, and standardize street abbreviations, zip codes, misspellings, and other address errors. In addition, users can add geo-coordinates to addresses. The software retains all of the raw address data and any changes or corrections made to the records, so users can access comprehensive address data for analysis.
MEASURING AND TUNING DATA

Once mapping and profiling are complete, the unified data sets are measured and tuned in preparation for grouping. In this stage, users quickly adjust matching threshold levels and set rules and exceptions for how and when to use common and unique data values during the grouping process. These capabilities are unique to Novetta Entity Analytics and were designed to eliminate much of the manual effort normally required to group data. Users can also easily filter out or apply automated processes to common data, which is not possible with other technologies.

First users select the strategy rules they want to enable and adjust the threshold setting the software will apply to each rule. A tight, moderate or loose threshold value can be chosen, depending on the number of false positive or false negative matches acceptable for the specific project (see Figure 7). Users who require more comprehensive threshold settings can enter them from the application’s developer interface.

After selecting rules and adjusting thresholds, users can view data uniqueness statistics and interactively set rule parameters to include or exclude specific data values within fields (see Figure 8).

ENTITY GROUPING AND RESOLUTION

Novetta Entity Analytics groups and resolves records into real-world entities by automatically applying a number of complex processes to the data. These processes partition data to increase processing speed and efficiencies, score and group attribute pairs, and resolve pairs into real-world entities.

The software’s highly accurate, built-in partitioning, scoring and resolution functions were developed by Novetta’s team of data and computer scientists as a result of their work with national security agencies and other U.S. government and commercial customers to combine and resolve thousands of data sources and hundreds of billions of entity records.
The unique distributed processing model included in Novetta Entity Analytics allows users to rapidly perform grouping operations on complete data sources, instead of samples, and immediately review results. This ability helps ensure the accuracy of entity records and prevents users from analyzing skewed or inaccurate data. The steps the software takes to automatically group and resolve entities include the following:

**Partition Data**

Novetta Entity Analytics applies built-in partitioning functions to loosely cluster records into small subsets of data likely to belong to the same entity. The partitioned data is efficiently distributed across a Hadoop cluster for rapid processing. Each strategy rule leverages different partitioning functions, based on the attributes included in the strategy. The software's overall goal is to create small partitions for faster processing, while avoiding over-partitioning of data and missing critical matches.

**Compare, Match and Score Records**

The software identifies matching pairs by automatically applying built-in scoring functions that perform comprehensive pair-wise comparisons of all records within each partition. By eliminating the need for users to write custom rules, these automated processes streamline the task of accurately grouping records. If custom rules are required, users can also add them from the application's developer interface.

**Build Resolved Entities**

Novetta Entity Analytics automatically recombines the partitioned data, builds real-world entities, based on the scores assigned to matched pairs, and assigns a unique entity ID to each entity. The software then assesses all of the data to detect any entity records that were erroneously assigned to multiple partitions and regroup them into a single real-world entity.

**Review Entities**

To determine if the required level of accuracy was achieved, users can review and verify either overall results, specific entity records, or a sampling of records from the initial resolution phase (see Figure 9).

If different results are desired, users can easily readjust threshold settings, uniqueness measures and scoring functions, and quickly run the entire grouping and entity resolution process again. Once optimum results are achieved, records can be published for immediate use or linking workflow processes can be applied to gain additional relationship insights from the data.

**ENTITY LINKING AND RELATIONSHIP ANALYSIS**

Linking processes quickly uncover nonobvious relationships within and between the real-world entities Novetta Entity Analytics has already resolved, or within other data sources. Built-in matching strategies, designed to detect simple relationships among and within entities, can be applied, e.g. associates who share the same phone number, address or organization, (see Figure 10). The software can also be customized to detect more complex relationships, such as temporal, behavioral and analytics links within and between entities.

---

*Figure 9. Decision Link Graph*

This graph displays a map of an individual entity and how all of their records are related to one another.
These links could include recent purchases, purchases made at a particular location at a specific time, or other behavioral patterns within transactional data.

The processes Novetta Entity Analytics uses to identify relationships within the data are similar to those used for grouping, but the strategy rules and partitioning functions are different. When partitioning data to detect relationships, the software uses more general attributes than those used for grouping, e.g., records with the same street name are partitioned together, instead of those with the same number and street name. Once pairs are scored within data partitions, links are produced that can be used in downstream processes to build graphs or enable more comprehensive analytics.

**PUBLISHING RESOLVED ENTITIES AND RELATIONSHIP RESULTS**

The final stage in the Novetta Entity Analytics grouping or linking workflow process is to publish entity and relationship data for use by other enterprise applications and internal workflows and processes. Data can be published as either labeled output or a multidimensional index.

A Hive table can be created that contains labeled output, including assigned entity IDs, attributes used by the strategy rules that grouped the entities, and, if desired, all or a subset of additional fields contained within the data. The output can then be published to HCatalog or exported for use by a variety of applications.

Output can also be published as a multidimensional index that contains entity IDs and source IDs. This output is often used by existing applications to create master IDs for real-world entities contained within their systems. Master IDs make it easier for users to find additional records about entities of interest within a system, e.g., if a query to an enterprise data warehouse returns a record about a specific entity, the user can easily look for additional records about that same entity within the system by searching on the entity ID.
ACCESSING ADDITIONAL CAPABILITIES

Users can access additional Novetta Entity Analytics capabilities through the software’s developer interface. The interface allows all features currently available from the application’s graphical user interface to be invoked through a RESTful API. In addition, other powerful capabilities are available from the application’s command-line interface, including the ability to:

- Cleanse and augment address information using address standardization functions, as discussed above.
- Identify entities within unstructured data using built-in annotated dictionaries of person names, organization names, organization types (e.g. Inc., LLC) street names and numbers, and other attributes, or by using custom dictionaries created from other structured or user-specified data sources.
- Improve overall entity accuracy by reviewing all records chained to each entity, applying built-in conflict resolution rules to identify potential resolution errors, and breaking apart entities or removing over-chained records from entities.
- Schedule and run jobs on a recurring basis, e.g. nightly, to ensure users are always presented with the latest data and that the data processed by Novetta Entity Analytics can easily fit within an automated data pipeline.

NOVETTA ENTITY ANALYTICS: DELIVERING POWERFUL ENTITY AND RELATIONSHIP INSIGHTS

Novetta Entity Analytics is a self-service data analytics platform that greatly improves efficiencies and delivers immediate value by enabling users to focus on their organization’s business needs instead of data preparation, management and manipulation. By following the steps described above, all user types, from data analysts to power users, can easily perform powerful entity resolution and relationship analysis on a wide range of data sources.
Headquartered in McLean, VA with over 700 employees across the US, Novetta has over two decades of experience solving problems of national significance through advanced analytics for government and commercial enterprises worldwide. Grounded in its work for national security clients, Novetta has pioneered disruptive technologies in four key areas of advanced analytics: data, cyber, open source/media and multi-int fusion. Novetta enables customers to find clarity from the complexity of ‘big data’ at the scale and speed needed to drive enterprise and mission success.