Data.com Record Matching in Salesforce

Salesforce, Spring '16

@salesforcedocs

Last updated: April 27, 2016

CONTENTS

DATA.COM RECORD MATCHING IN SALESFORCE

Introduction

At Salesforce, our goal is to provide Data.com account, contact, and lead data that starts with and maintains the highest quality at all times. Based on input from our customers, we measure this new standard of data quality against four key metrics.

- **Accuracy**–We start with leading industry data providers and a two-million member passionate community of data contributors that provide and monitor content for accuracy, supported by innovative technologies and internal stewardship operations.
- **Completeness.**–We provide the data you need to identify ideal accounts, plan territories, identify leads, and connect quickly by phone or email.
- **Freshness.**–Our data is managed in the cloud, so we can continuously obtain and provide updates to our customers.
- **Coverage.**–We provide easy, integrated access to the world's most robust B2B database, which connects you with more companies—and more contacts at every level of every account.

The ability to maintain high quality Salesforce account, contact, and lead record data is a key component of Salesforce's Data.com product suite. In addition to sales prospecting, the Data.com Corporate and Data.com Premium products offer manual cleaning for all licensed users. The Data.com Clean product, which can be purchased for use on its own or together with Corporate or Premium, provides manual clean functionality as well as automated jobs for cleaning account, contact, and lead records.

To *clean* a record is to bring it up to date with the values you want—either accepting some or all of the latest values from Data.com, or keeping all of the values in the Salesforce record. Record matching is the foundation of Data.com's Clean functionality. This paper explains how matching works.

Matching Process Overview

Within Salesforce, the matching process for account, contact, and lead records can be initiated in one of two ways.

- Manually, when a user clicks the **Clean** button on an individual account, contact, or lead record.
- Via clean jobs, which attempt to match all your account, contact, or lead records automatically and either flag different values or auto-fill fields with Data.com values.

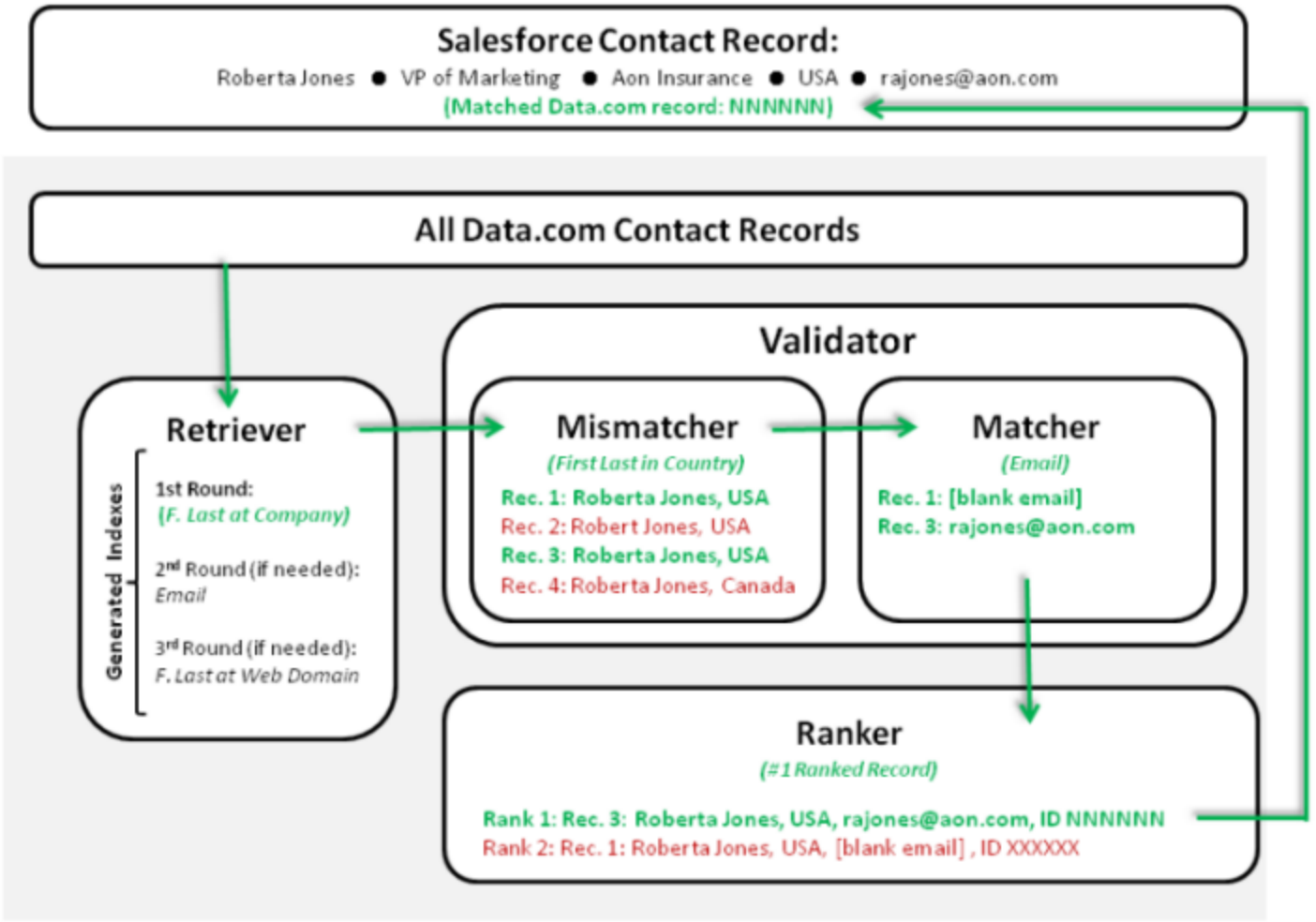
Note: *Automated* clean jobs are only available to organizations that purchase Data.com Clean. *Manual* clean is available to *all* licensed Data.com Corporate and Premium users.

The first time a Salesforce record is cleaned—either manually or via an automated job—the matching algorithm goes to work. It searches Data.com for similar records, examines the field values in each record, and looks for enough identical or similar information to indicate that the two records represent the same entity (that is—the same account, contact, or lead). If the algorithm finds a match in Data.com for the Salesforce record, it links the two records by placing a numeric value in their respective `Data.com Key` fields. This value represents the current mapping, which is maintained when the record is cleaned in the future. If the Salesforce record changes substantially, the next time it is cleaned (either manually or via jobs), the matching engine may look for a better match from other Data.com records. If a better match is found, a new `Data.com Key` field value may replace the current one. If more than one potential match is found in Data.com, identical `Data.com Key` field values are the tie breaker.

The Matching Algorithm in Detail

Data.com's matching algorithm uses "fuzzy" matching, which considers the values in key fields and sets of fields when attempting to match account, contact, and lead records. Thanks to fuzzy matching, it's not necessary for all of the Salesforce record's fields to be populated. Nor is it necessary that every value be 100% accurate, but a match is more likely to occur if a field has *no* value than if it has a *conflicting* value.

This section explains the Data.com matching algorithm and its component subsystems. For information on getting better match rates in Salesforce, see "Improving your Data.com Match Rates," below.



Indexer / Retriever

Data.com stores records in a cache, which contains a record index that we rebuild several times a year. Each time we rebuild the index, we normalize the field values within the records, transforming or replacing values if necessary. For example, `Rob` and `Bob` might be normalized to `Robert`, and `Coke` and `The Coca-Cola Company` might be normalized to `Coca-Cola`. We also look for the existence of values in certain meaningful Data.com fields (such as `First Name`, `Company Domain`, and `City`) and use these

fields either alone or in combinations to create *index entries*, such as `flast@company` (which stands for first initial plus last name combined with company).

If the cache contains a number of records that after field normalization have the same values for an index entry, the indexer derives a *key* for that set of records. Each key, itself, has a value, which is the list of IDs of Data.com records that match that key.

When a Salesforce (input) record is cleaned, we use the same process of normalizing and combining field values to create a key for that record. We use the Salesforce record's key to search the Data.com index for a matching key, and if we find one, we retrieve the Data.com records in that key from the cache. Those records become match candidates for the Salesforce record. Now it's time to validate them.

Validator: Mismatcher

The algorithm's Mismatch Identifier (Mismatcher) is the first stage in the Validator process. It takes the Retriever's list of Data.com match candidate records and works through them one by one. For each type of record (account, contact, or lead), Mismatcher examines the values in the candidate and input record for a small, predefined set of indicator fields, including `First Name` for contacts and `Country` for accounts. If the indicator field values in the input record and a match candidate record are sufficiently different, the Data.com record will be rejected.

Validator: Matcher

The algorithm's Match Identifier (Matcher) is the second stage in the Validator process. Its job is to identify potential matches among the remaining match candidate records: those that have passed the Mismatcher. For each record type (account, contact, or lead), the Matcher uses much wider sets of indicator fields, including `Last Name`, `Company Name`, and `Title` for contacts, and `Web Domain` and `Address` for accounts.

For contacts, Matcher employs a department-compatibility matrix to identify department names that while technically different, might mean the same thing. For example, the title **Marketing Director** might be considered similar enough to the title **Director of Strategic Marketing** for the two records to be matched (if other field values meet matching requirements, as well). Matcher also employs probability matching, which combines the comparative commonness of a person's name with the size of the company where they work to create matches. For example, two contacts named "Kevin Akeyroyd" at salesforce.com are likely to be duplicates, because even though salesforce.com is a comparatively large company, the last name Akeroyd is not common enough for more than one Kevin with that last name to work there. By contrast, two Nancy Smiths at Smith Saddlery are likely to be duplicates because Smith Saddlery is a small company.

Note that the algorithm does not use probability matching for accounts.

If the values in the input record and a match candidate record are sufficiently different, the Data.com record will not meet the algorithm's predefined threshold, and the Data.com record will be rejected---even if the two records' other field values are identical. If all records are rejected at any point in the Validator process (Mismatcher or Matcher), the entire matching process starts again: the Retriever creates a new key from the input record and searches for a matching key within the Data.com record cache.

Ranker

Ranker sorts the records it receives from the Matcher and returns the them to the API, with the matching record ranked first. That record and the Salesforce record are linked in Salesforce and will remain linked until one or both records change substantially.

Improving Your Match Rates

Data.com's matching algorithm is designed to find the best matches as often as possible while avoiding false matches. From our analysis of customer data, we've found that match rates of around 40 to 50 percent are typical, but results can vary.

For the best chance at matching with Data.com records, it's important that your Salesforce records have accurate and complete values for certain important fields.

For contacts and leads, the important fields are:

- `Name`
- `Email`

  Ideally, use a direct email address rather than a group address like info@org.com. (Identical Email values will almost always trigger a match unless other values in the record conflict.) **Note:** Data.com does not store emails that may be personal contact information, such as those from Gmail or Yahoo.

- `Account Name` for contacts; `Company` for leads.
- `Title`
- `Phone`

For accounts, the important fields are:

- `Account Name`
  - Make sure the account name doesn't contain any unrelated artifacts, such as numbers (1002), special characters (!#@#), or unrelated words. (These symbols are acceptable if they are part of the company's name.)
  - Try to avoid country and state names as values unless they are part of the company's name.
  - If the account name contains more than one word, like *`DSGI Business (PC World Business)`* try both of them.
  - If an account name is also a Website (like *`salesforce.com`*), try leaving off the subdomain (*`.com`* and the like).
- `Billing Address`
  - Use a full address if possible
  - Make sure the `Country` field value is correct. If you don't know the country, leave the field blank.
  - Make sure to specify the `State` or `ZIP code` value if you know it.
  - If you know the account's street name but not its street number, use the street name only. A street name without a street number is better than no value.
- `Website`
- `Phone`

Matching Examples

Here are some examples (both detailed and basic) that illustrate matching outcomes.

Contacts

1: Salesforce record with partial data:

First Name	Last Name	Account Name	Title	Phone	Email
Parker	*Harris*	*salesforce.com*			*parkerh@salesforce.com*

Data.com record match candidates:

Rank	First Name	Last Name	Company Name	Title	Phone	Email
1	*Parker*	*Harris*	*salesforce.com*	*Executive Vice President, Technology, salesforce.com*	*+1.415.901.7000*	*parkerh@salesforce.com*
2	*Parker*	*Harris*	*salesforce.com*	*Executive Vice President, Technology and Products*	*+1.415.901.7021*	*parkerh@salesforce.com*

2: Salesforce record with out-of-date information:

First Name	Last Name	Account Name	Title	Phone	Email
Brett	*Queener*	*salesforce.com*	*VP, Field Ops*		

Data.com record match candidates:

Rank	First Name	Last Name	Company Name	Title	Phone	Email
1	*Brett*	*Queener*	*salesforce.com*	*Executive Vice President and General Manager, Data.com*	*+1.415.901.8473*	*bqueener@salesforce.com*
2	*Brett*	*Queener*	*salesforce.com*	*Vice President, Field Operations*	*+1.415.901.8473*	*bqueener@salesforce.com*
3	*Brett*	*Queener*	*salesforce.com*	*Executive Vice President, Applications*	*+1.415.901.8473*	*bqueener@salesforce.com*

Accounts

1: Salesforce record with complete data:

Account Name	Phone	Website	Billing Street	Billing City	Billing ZIP	Billing Country
Google Inc.	*+1.650.253.0000*	*www.google.com*	*1600 Ampitheatre Parkway*	*Mountain View*	*94043*	*USA*

Data.com record match candidates. Matches headquarters address.

Rank	Company Name	Phone	Website	Billing Street	Billing City	Billing ZIP	Billing Country
1	*Google Inc.*	*+1.650.253.0000*	*www.google.com*	*1600 Ampitheatre Parkway*	*Mountain View*	*94043*	*United States*
2	*Google, Inc.*		*www.google.com*	*1600 Ampitheatre Pkwy Building #42*	*Mountain View*	*94043-1351*	*United States*
NA	*Aladdin Client Google*	*+1.650.967.3916*	*www.google.com*	*1400 Crittenden Ln*	*Mountain View*	*94043-2775*	*United States*
NA	*Google Inc.*		*www.google.com*		*Mountain View*	*94042*	*United States*
NA	*Google Inc*	*+1.650.253.1525*	*www.google.com*	*2690 Casey Ave.*	*Mountain View*	*94043-1141*	*United States*
NA	*Google, Inc.*	*+1.650.214.3050*	*www.google.com*	*2350 Bayshore Pkwy*	*Mountain View*	*94043-1121*	*United States*

2: Salesforce record with no `Billing Street` value:

Account Name	Phone	Website	Billing Street	Billing City	Billing ZIP	Billing Country
Google Inc.		*www.google.com*		*Mountain View*		

Data.com record match candidates. Matches HQ address with highest contact count (not shown).

Rank	Company Name	Phone	Website	Billing Street	Billing City	Billing ZIP	Billing Country
1	*Google Inc.*	*+1.650. 253.0000*	*www.google.com*	*1600 Ampitheatre Parkway*	*Mountain View*	*94043*	*United States*
NA	*Aladdin Client Google*	*+1.650. 967.3916*	*www.google.com*	*1400 Crittenden Ln*	*Mountain View*	*94043-2775*	*United States*
3	*Google, Inc.*		*www.google.com*	*1600 Ampitheatre Pkwy Building #42*	*Mountain View*	*94043-1351*	*United States*
4	*Google Inc.*		*www.google.com*		*Mountain View*	*94042*	*United States*
5	*Google Inc*	*+1.650. 253.1525*	*www.google.com*	*2690 Casey Ave.*	*Mountain View*	*94043-1141*	*United States*
6	*Google, Inc.*	*+1.650. 214.3050*	*www.google.com*	*2350 Bayshore Pkwy*	*Mountain View*	*94043-1121*	*United States*

3: Salesforce record with partial `Billing Street` value:

Account Name	Phone	Website	Billing Street	Billing City	Billing ZIP	Billing Country
Google Inc.		*www.google.com*	*Bayshore*	*Mountain View*		

Data.com record match candidates. Matches HQ address with highest contact count (not shown).

Rank	Company Name	Phone	Website	Billing Street	Billing City	Billing ZIP	Billing Country
1	*Google International LLC*	*+1.650. 643.4000*	*www.google.com*	*2400 Bayshore Parkway*	*Mountain View*	*94043-1103*	*United States*
2	*Google, Inc.*	*+1.650. 214.3050*	*www.google.com*	*2350 Bayshore Pkwy*	*Mountain View*	*94043-1121*	*United States*
NA	*Google, Inc.*		*www.google.com*	*1600 Ampitheatre Pkwy*	*Mountain View*	*94043-1351*	*United States*

Rank	Company Name	Phone	Website	Billing Street	Billing City	Billing ZIP	Billing Country
				Building #42			
NA	*Aladdin Client Google*	*+1.650. 967.3916*	*www.google.com*	*1400 Crittenden Ln*	*Mountain View*	*94043-2775*	*United States*
NA	*Google Inc.*	*+1.650. 253.0000*	*www.google.com*	*1600 Ampitheatre Parkway*	*Mountain View*	*94043-1351*	*United States*
NA	*Google Inc.*		*www.google.com*		*Mountain View*	*94042*	*United States*
NA	*Google Inc*	*+1.650. 253.1525*	*www.google.com*	*2690 Casey Ave.*	*Mountain View*	*94043-1141*	*United States*
NA	*Google, Inc.*	*+1.253.3749*	*www.google.com*	*1945 Charleston Rd*	*Mountain View*	*94043-1201*	*United States*