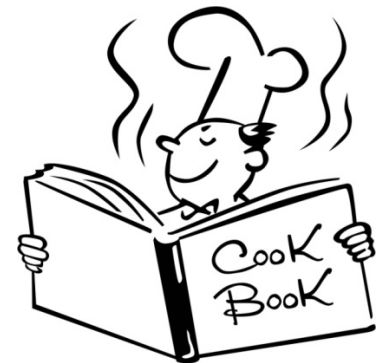


# Company Name Cleaning & Normalization

**OPEN**RISE

Cook Book Series



# Recipe Overview

Build Company Master List

Run automatically

Rule 1

CoMaster R1 -  
Clean Company  
Name

Rule 2

CoMaster R2 -  
Build Master List

Norm Company Name To Master

Run automatically

Rule 1

NormCo R1 -  
Norm Company To  
Master

This is a recipe for cleaning and normalizing company name data

- Clean and reformat company names for readability
- Create company-alias master list
- Normalize company name data using master list

You will need the following raw data:

- Company name

# tips

- Add a rule by clicking on an existing rule  and +.

- Put new data into a new data attribute so you can easily compare before vs. after and confirm the rule is doing what it is supposed to do.

- Some configurations are found by clicking on: 

- Can't see the open reference data? Check the setting in your Data Catalog:



- The company-alias master list is generated using a machine algorithm. It is very accurate but never perfect. It is highly recommended that you review and tweak the master list before using it to normalize company names.
- Experiment with the fuzzy matching algorithm parameters to get the best results.

# Pipeline 1, Rule 1: Clean Company Name

**Rule name \***  
CoMaster R1 - Clean Company Name

**Rule template \* Need help picking a template?**  
Company name clean up

**IF this happens**

**Input Data Sources**  
Marketo 0815

**SELECT DATA**

All new data

TIP: Options to remove or expand words like Inc, Corp, Ltd. We highly recommend removal. Easy to read and generates better master list.

## Reference Data



**Reference - Company Name Clean Up**  
List of common company name extensions used to clean up company names. Extensions such as Inc, Corp, Ltd, LLP, Pvt, AB, SAB, NV, .... etc.

**THEN take these actions**

**Output Data Sources**  
Automatically created after rule is saved.

**Clean up company name**

Clean up the company name in: company

Write cleaned name to: company\_clean

**ADD ATTRIBUTE**

Format to use for cleaned name: Remove words like Inc, Corp, LLC

**Advanced configurations**

Company keywords are stored in: Reference - Company Name Clean Up

Normalized values are stored in: Complete Word

Alias values are stored in: Company Name Keywords

Show less

## Cleaned Company Names With Inc, Corp, Ltd **Removed**

company	company_clean
Symantec Corporation	Symantec
Nomura Research Institute America, Inc.	Nomura Research Institute America
Microsoft Corporation	Microsoft
MTG Management Consultants LLC	Mtg Management Consultants
McAfee, Inc.	Mcafee
EMC Corporation	Emc
Symantec Corporation	Symantec
Fiserv, Inc.	Fiserv
DigitalPersona, Inc.	Digitalpersona
Concurrent Technologies Corporation	Concurrent Technologies
Honeywell International Inc.	Honeywell International
Lab Spin Inc.	Lab Spin
Symantec Corporation	Symantec
Qualys, Inc.	Qualys
Symantec Corporation	Symantec
EMC Corporation	Emc
Cenovus Energy Inc.	Cenovus Energy
403 Labs, LLC	403 Labs
EMC Corporation	Emc
EMC Corporation	Emc
Health IT Law Group LLC	Health It Law Group
Time Warner Inc.	Time Warner
Zix Corporation	Zix
AT&T Inc.	At&t
Webroot, Inc.	Webroot
TELUS Communications Inc	Telus Communications
MITRE Corporation	Mitre
EMC Corporation	Emc
Qualys, Inc.	Qualys

## Cleaned Company Names With Inc, Corp, Ltd **Expanded**

company	company_clean
MTG Management Consultants LLC	Mtg Management Consultants LLC
paycorp	Paycorp
Mindtree Ltd	Mindtree Limited
McAfee, Inc.	Mcafee Incorporated
EMC Corporation	Emc Corporation
Aquad Inc.	Aquad Incorporated
McAfee, Inc.	Mcafee Incorporated
EMC Corporation	Emc Corporation
Fiserv, Inc.	Fiserv Incorporated
Aquion Pty Ltd	Aquion Pty Limited
Qualys, Inc.	Qualys Incorporated
ST Electronics (Info-security) Pte Ltd	St Electronics Info-security Private Limited
Qualys, Inc.	Qualys Incorporated
TELUS Communications Inc	Telus Communications Incorporated
Symantec Corporation	Symantec Corporation
EMC Corporation	Emc Corporation
403 Labs, LLC	403 Labs LLC
EMC Corporation	Emc Corporation
Microsoft Corporation	Microsoft Corporation
TELUS Communications Inc	Telus Communications Incorporated
Time Warner Inc.	Time Warner Incorporated
NSFOCUS Information Technology Co.,Ltd.	Nsfocus Information Technology Company Limited
Imperva Inc.	Imperva Incorporated
AT&T Inc.	At&t Incorporated
Health IT Law Group LLC	Health It Law Group LLC
Synopsys, Inc.	Synopsys Incorporated
Webroot, Inc.	Webroot Incorporated
VeriSign, Inc.	Verisign Incorporated
TELUS Communications Inc	Telus Communications Incorporated

# Pipeline 1, Rule 2: Build Master List

**Rule name \***  
CoMaster R2 - Build Master List

**Rule template \*** [Need help picking a template?](#)  
Generate normalization reference

**IF this happens**

**Input Data Sources**  
Rule Output - CoMaster R1 - Clean Company Name

**SELECT DATA**

All new data

Make sure to use the cleaned company name, not the original company name.

## THEN take these actions

### Output Data Sources

Automatically created after rule is saved.

### Generate normalization reference

Generate normalization reference data for this attributes:

company\_clean

### Advanced configurations

Degree of fuzziness

(0.1 to 1, 0.1 is maximum fuzziness)

0.1

Percentage of leading text that must match

60

Ignore if characters less than

3

[Show less](#)

Start with these default values. See tuning tips on the next page.

## Company-Alias Master List Generated

company_clean	company_clean Alias
At&t	atat,it&t,at&t nj,atnt,at&t
Hsbc	hsbc bank north america,hsbc,hsbc holdings
Dell	dell,dell sonicwall,dell boomi,dell services
Adobe	adobe systems,adobe
Orange	orange,orange poland,orange regional medical center,orange sv,orange busines services,orange france telecom,orange slovakia
Toyota	toyota new zealand,toyota motors sales usa,toyota motor sales usa,toyota motor sales usa in,toyota usa,toyota motor australia,toyota canada,toyota financial services,toyota
Nissan	nissan north america,nissan
Siemens	siemens north america,siemens siemens enterprise communications gmbh,siemens osram opto semiconductors & ohg,siemens cinterion wireless modules,siemens siemens medical solutions,siemens loher,siemens medical solutions usa,siemens tochter,siemens siemens financial services,siemens siemens it solutions & services,siemens bfe studio und medien systeme,siemens evosoft,siemens product lifecycle management software de,siemens it solutions and services information technology,siemens,siemens home and office communication devices,siemens fujitsu technology solutions,siemens energy,siemens enterprise communications
Verizon	verizon wireless,verizon wireline,verizon,verizion,verizon business
Vodafone	vodafone group,vodafone,vodafone nz,vodafone qatar
Fidelity	fidelity,fidelity investment,fidelity international,fidelity investments
Net Orange	net orange
Wells Fargo	wells fargo bank na,wells fargo

# tips

- The higher the fuzziness index, the more closely the names have to match to be grouped together. For example:
  - “UBS Financial” and “ABC Financial” will match on high index  $\sim 0.8$
  - “UBS Financial” and “UBC Finland” will match on lower index  $\sim 0.3$

Degree of fuzziness

(0.1 to 1, 0.1 is maximum fuzziness)

0.1

- The leading index dictates what % of leading text must match for the names to be grouped together. For example:
  - “Department of Motor Vehicles Arizona” and “Department of Motor Vehicles Alabama” will match on an index of 70%
  - “DMV Arizona and DMV Alabama” will not match on an index of 70%

Percentage of leading text that must match

60

- Short names can create many false groupings. Increase minimum character index to reduce matching on short names. For example:
  - CSC vs. USC, or NBC vs. NBA

Ignore if characters less than

3



# Pipeline 2, Rule 1: Normalize Co. Name

**Rule name \***  
NormCo R1 - Norm Company To Master

**Rule template \*** [Need help picking a template?](#)  
Simple replacement / normalization

**IF this happens**

**Input Data Sources**  
Rule Output - CoMaster R1 - Clean Company Name

**SELECT DATA**

Normalize the cleaned company names produced by Pipeline 1 Rule 1

**THEN take these actions**

**Output Data Sources**  
Rule Output - NormCo R1 - Norm Company To Master

**Set attribute value**

Attribute to normalize / clean up :	company_clean
Write normalized value to	company_clean_norm
	<b>ADD ATTRIBUTE</b>
Normalized-alias mapping is stored in	Rule Output - CoMaster R2 - Build
Normalized value is stored in	company_clean
Alias is stored in	company_clean Alias
Match method	Fuzzy match
Resolve conflicts using the priority in	- None -

Reference is the Master List produced by Pipeline 1 Rule 2

## Company Names Cleaned Then Normalized

company	company_clean	company_norm_clean
Toyota Motor Sales, U.S.A., Inc.	Toyota Motor Sales Usa	Toyota
TOYOTA MOTOR SALES USA INC	Toyota Motor Sales Usa	Toyota
TOYOTA MOTOR SALES USA IN	Toyota Motor Sales Usa In	Toyota
TOYOTA MOTOR SALES U.S.A. INC.	Toyota Motor Sales Usa	Toyota
TOYOTA MOTOR SALES USA INC	Toyota Motor Sales Usa	Toyota
TOYOTA MOTOR SALES USA INC	Toyota Motor Sales Usa	Toyota
TOYOTA MOTOR SALES USA INCORPORATED	Toyota Motor Sales Usa	Toyota
TOYOTA MOTOR SALES USA INC	Toyota Motor Sales Usa	Toyota
TOYOTA MOTOR SALES USA INC	Toyota Motor Sales Usa	Toyota
TOYOTA MOTOR SALES USA INC	Toyota Motor Sales Usa	Toyota
TOYOTA MOTOR SALES USA INC	Toyota Motor Sales Usa	Toyota
TOYOTA MOTOR SALES USA INC.	Toyota Motor Sales Usa	Toyota
TOYOTA MOTOR SALES USA INCORPORATED	Toyota Motor Sales Usa	Toyota
TOYOTA MOTOR SALES USA INC	Toyota Motor Sales Usa	Toyota
TOYOTA MOTOR SALES U.S.A. INC.	Toyota Motor Sales Usa	Toyota
TOYOTA MOTOR SALES USA INC	Toyota Motor Sales Usa	Toyota
TOYOTA MOTOR SALES USA	Toyota Motor Sales Usa	Toyota
TOYOTA MOTOR SALES USA IN	Toyota Motor Sales Usa In	Toyota
TOYOTA MOTOR SALES U S A INC	Toyota Motor Sales Usa	Toyota
TOYOTA MOTOR SALES USA INC	Toyota Motor Sales Usa	Toyota
TOYOTA MOTOR SALES USA	Toyota Motor Sales Usa	Toyota
TOYOTA MOTOR SALES USA INC	Toyota Motor Sales Usa	Toyota
Toyota USA	Toyota Usa	Toyota

# Recipe Review

Build Company Master List

Run automatically

Rule 1

CoMaster R1 -  
Clean Company  
Name

Rule 2

CoMaster R2 -  
Build Master List

Norm Company Name To Master

Run automatically

Rule 1

NormCo R1 -  
Norm Company To  
Master

## Recommendations

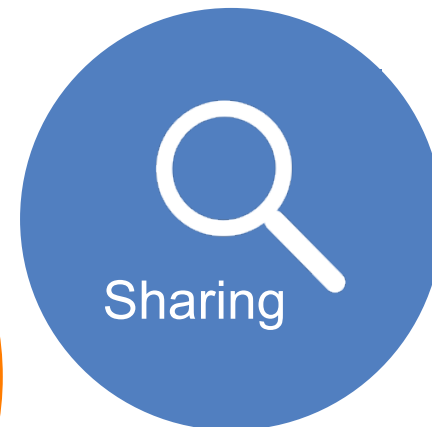
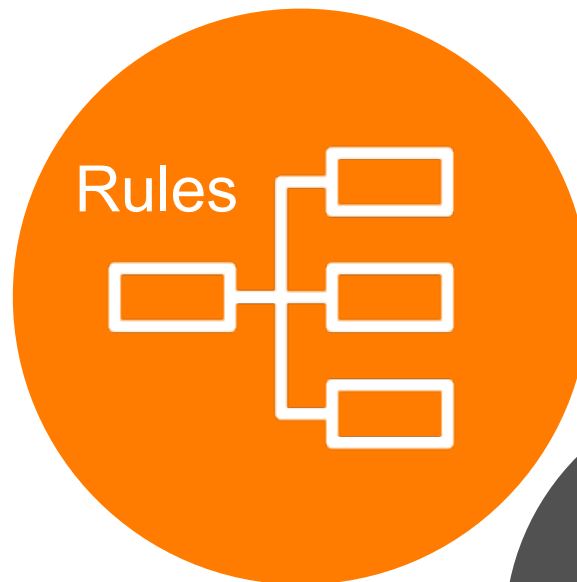
- For marketing systems, consider reducing the master list down to only customers and target accounts. It greatly reduces maintenance efforts.

Want to do more? Try the following on your own:

- In addition to normalizing company name, add parent company data to the master list and append pipeline and sales data with parent company information. This enables aggregated reporting and account based marketing.

# OPENPRISE

Data Automation For Business Users



[info@openprisetech.com](mailto:info@openprisetech.com)

Twitter: @openprisetech

[www.openprisetech.com](http://www.openprisetech.com)