

Configuring Natural Language Search

Date published: 2020-10-30

Date modified: 2024-05-30



Legal Notice

© Cloudera Inc. 2024. All rights reserved.

The documentation is and contains Cloudera proprietary information protected by copyright and other intellectual property rights. No license under copyright or any other intellectual property right is granted herein.

Unless otherwise noted, scripts and sample code are licensed under the Apache License, Version 2.0.

Copyright information for Cloudera software may be found within the documentation accompanying each component in a particular release.

Cloudera software includes software from various open source or other third party projects, and may be released under the Apache Software License 2.0 (“ASLv2”), the Affero General Public License version 3 (AGPLv3), or other license terms. Other software included may be released under the terms of alternative open source licenses. Please review the license and notice files accompanying the software for additional licensing information.

Please visit the Cloudera software product page for more information on Cloudera software. For more information on Cloudera support services, please visit either the Support or Sales page. Feel free to contact us directly to discuss your specific needs.

Cloudera reserves the right to change any products at any time, and without notice. Cloudera assumes no responsibility nor liability arising from the use of products, except as expressly agreed to in writing by Cloudera.

Cloudera, Cloudera Altus, HUE, Impala, Cloudera Impala, and other Cloudera marks are registered or unregistered trademarks in the United States and other countries. All other trademarks are the property of their respective owners.

Disclaimer: EXCEPT AS EXPRESSLY PROVIDED IN A WRITTEN AGREEMENT WITH CLOUDERA, CLOUDERA DOES NOT MAKE NOR GIVE ANY REPRESENTATION, WARRANTY, NOR COVENANT OF ANY KIND, WHETHER EXPRESS OR IMPLIED, IN CONNECTION WITH CLOUDERA TECHNOLOGY OR RELATED SUPPORT PROVIDED IN CONNECTION THEREWITH. CLOUDERA DOES NOT WARRANT THAT CLOUDERA PRODUCTS NOR SOFTWARE WILL OPERATE UNINTERRUPTED NOR THAT IT WILL BE FREE FROM DEFECTS NOR ERRORS, THAT IT WILL PROTECT YOUR DATA FROM LOSS, CORRUPTION NOR UNAVAILABILITY, NOR THAT IT WILL MEET ALL OF CUSTOMER’S BUSINESS REQUIREMENTS. WITHOUT LIMITING THE FOREGOING, AND TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, CLOUDERA EXPRESSLY DISCLAIMS ANY AND ALL IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, QUALITY, NON-INFRINGEMENT, TITLE, AND FITNESS FOR A PARTICULAR PURPOSE AND ANY REPRESENTATION, WARRANTY, OR COVENANT BASED ON COURSE OF DEALING OR USAGE IN TRADE.

Contents

Enabling Natural Language Search in Site Settings.....	4
Enabling a dataset for NLS.....	5
Specifying a group for a dataset in NLS.....	7
Specifying fields for a dataset in NLS.....	8
Specifying synonyms for a dataset in NLS.....	9
Specifying default aggregation field for a dataset in NLS.....	10
Specifying suggested questions for a dataset in NLS.....	13
Specifying word substitutions for a dataset in NLS.....	14
Configuring date and time for search requirements.....	15
Specifying the date/time field and the format of date/time for dataset.....	16
Specifying the default time interval in dataset.....	17

Enabling Natural Language Search in Site Settings

Before using Natural Language Search (NLS) in Data Visualization, you must enable the feature in the Site Settings.

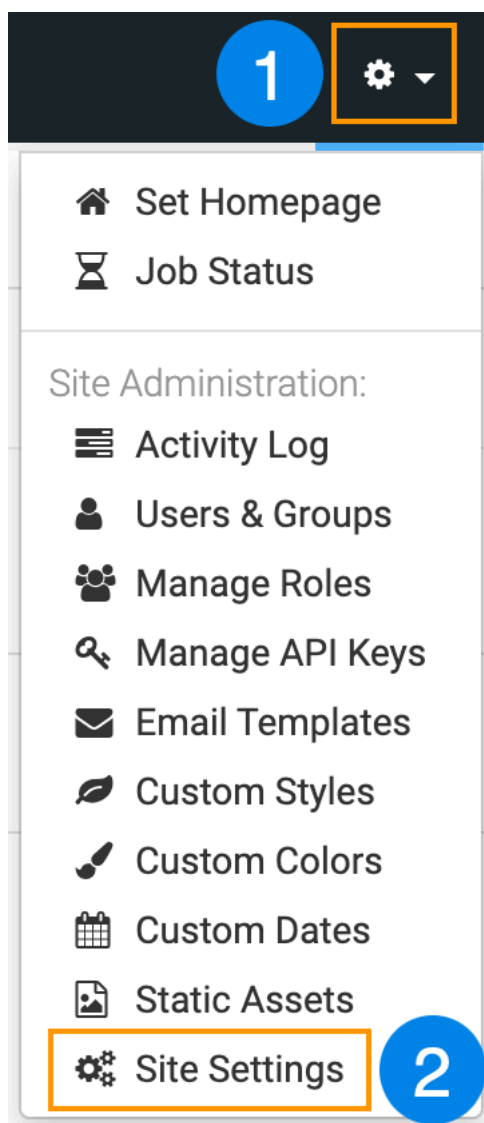
About this task



Note: This setting is only available for users with administrative privileges.

Procedure

1. Click the Gear icon in the upper right corner to open the Site Administration menu.
2. Click Site Settings.



The settings page appears.

3. Select the following options under Features:

- Enable Search - You have to enable this option for the other two to take effect.
- Enable Search in Menu Bar - This option allows you to open the search modal on any screen from the icon from the Search button in the top right corner.
- Enable Search Visual - This option allows you to add search visuals to a dashboard.

Features

- ☒ Enable Derived Data
- ☒ Enable Custom Styles
- ☒ Enable Data Downloads from URL
- ☒ Enable Search (Tech Preview)
- ☒ Enable Search in Menu Bar (Tech Preview) ⓘ
- ☒ Enable Search Visual (Tech Preview) ⓘ

4. Click SAVE in the upper left corner.

Enabling a dataset for NLS

Before configuring Natural Language Search (NLS) in Data Visualization, you must enable and configure the datasets.

Procedure

1. On the main navigation bar, click DATA.



2. In the list of datasets, click the one you want to enable for NLS.

The Dataset Detail page appears.

In the figure below, the Restaurant Inspection SF dataset is selected.

The screenshot shows the Cloudera Data Visualization interface. The top navigation bar includes 'HOME', 'VISUALS', and 'DATA'. The left sidebar contains a menu with 'Dataset Detail', 'Related Dashboards' (1), 'Fields', 'Data Model', 'Time Modeling', 'Search Modeling' (highlighted with an orange box), 'Segments' (0), 'Filter Associations' (0), 'Permissions', and 'Extract Job Logs'. The main content area is titled 'Dataset: Restaurant Inspection SF' and 'Detail'. It lists the following information: Dataset: Restaurant Inspection SF (with an edit icon), Table: main.restaurant_scores_lives_standard, Connection Type: SQLite, Data Connection: samples (with an edit icon), Description: (with an edit icon), Join Elimination: Enabled (with an edit icon), and Result Cache: From Connection (with edit and refresh icons). Below this, it shows ID: 5, Created on: May 26, 2021 01:55 PM, Created by: (redacted), Last updated: Jun 13, 2021 01:55 PM, and Last updated by: (redacted).

3. Click Search Modeling.
4. In the Search Modeling page, select Enable search on this dataset.

The screenshot shows the 'Search Modeling' page for the 'Restaurant Inspection SF' dataset. The page title is 'Dataset: Restaurant Inspection SF' and 'Search Modeling'. There are two buttons: 'UNDO' and 'SAVE'. Below these, there is a checkbox labeled 'Enable search on this dataset' which is checked and highlighted with an orange box.

5. Click SAVE.



Note: On the list of datasets, a search-enabled dataset has an active Search icon.



Note: To disable the search on a dataset, clear the Enable search on this dataset checkbox.

Specifying a group for a dataset in NLS

Data Visualization enables you to group datasets for Natural Language Search (NLS). You can use a group to more easily combine datasets during search operations.

Procedure

1. Under Search Group, add a new name for the group or choose one from the already existing ones.

In the figure below, the dataset is added to the group 'Inspections'.

Dataset: Restaurant Inspection SF

Search Modeling

UNDO SAVE

☒ Enable search on this dataset

Search Group

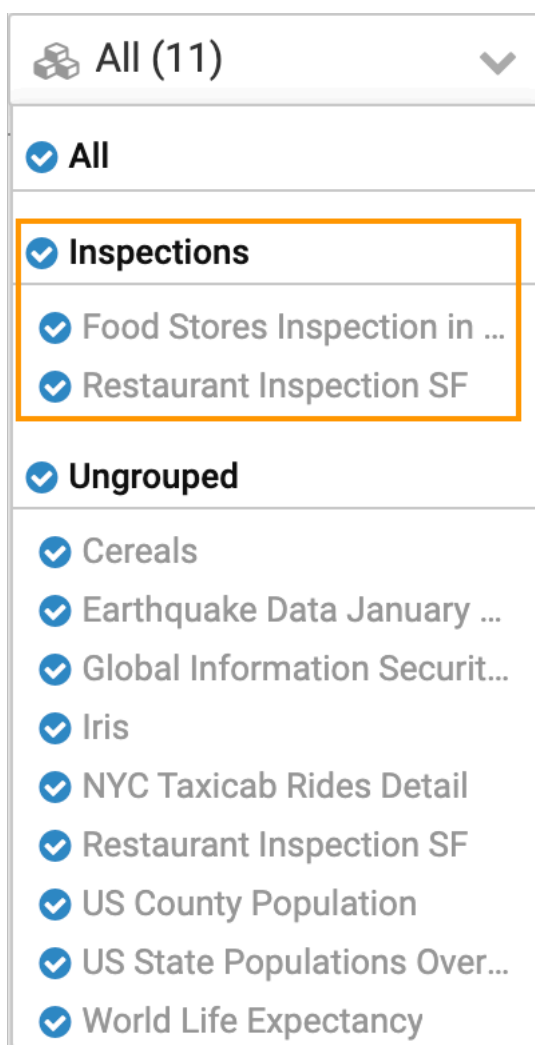
Quickly find answers in all datasets that belong to the same search group.

Inspections

2. Click SAVE.

Results

You can see the datasets grouped in the drop-down list in the Search modal window.



Specifying fields for a dataset in NLS

Data Visualization makes it possible to limit Natural Language Search (NLS) to values of specified fields of the dataset.

About this task

You can restrict the search for a number of reasons, either from operational considerations, or for security reasons.

Procedure

1. Under Fields, specify which fields may be searched through NLS.

In the example below, most of the fields are selected. If you want to disable search for a specific field in a dataset, simply remove all words and phrases from Matching search terms of that field. In the example below, the terms from Record Count were removed.

- Choose an option from the drop-down list under Auto Search Priority.

Higher priority columns are searched first, then medium, and finally low.



Note: Enabling Auto Search Priority will result in additional queries even if the column is not specifically mentioned in the search and can affect performance. Only enable it for columns from which users want to pull information.

Fields

Field	Data Type	Dimension or Aggregate	Matching search terms ⓘ	Auto Search Priority ⓘ
[Record Count]	BIGINT	Aggregate		NONE ▾
[business_id]	BIGINT	Dimension	business_id	HIGH ▾
[business_name]	STRING	Dimension	business establishment restaurant	LOW ▾
[business_address]	STRING	Dimension	business_address	NONE ▾
[business_city]	STRING	Dimension	business_city city	HIGH ▾
[business_state]	STRING	Dimension	business_state state	MEDIUM ▾
[business_postal_code]	STRING	Dimension	"postal code" zip	LOW ▾
[business_latitude]	DOUBLE	Dimension	"business latitude" lat	LOW ▾

- Click SAVE.

Specifying synonyms for a dataset in NLS

With Data Visualization, you can specify synonyms for dataset fields; these words can be used in Natural Language Search (NLS).

Procedure

- To specify alternate words for the dataset fields, you can add them to the column Matching search terms. Separate synonyms by including a single space, and use quotations for phrases.
- Under Fields, add synonymous terms that can be searched through NLS.

In the example below the terms 'business name', 'establishment', 'business', 'restaurant', and 'cafe' are added for the column 'Business Name'.

Fields

Field	Data Type	Dimension or Aggregate	Matching search terms ⓘ	Auto Search Priority ⓘ
[Record Count]	BIGINT	Aggregate		NONE ▾
[business_id]	BIGINT	Dimension	business_id	NONE ▾
[business_name]	STRING	Dimension	"business name" establishment business restaurant cafe	LOW ▾

- Click SAVE.

Specifying default aggregation field for a dataset in NLS

In Data Visualization, you can specify which field's values report the default aggregation.

About this task

When a search phrase does not specify an aggregate measure, Data Visualization reports results of a default column as an aggregate. In the majority of cases, this default column is Record Count, a default aggregation measurement generated by the system.

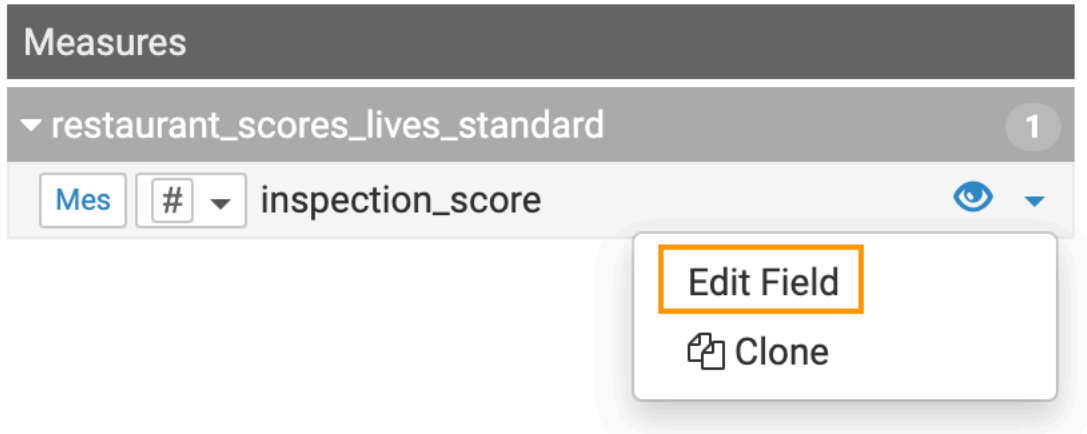
In this example, `inspection_score` is selected to illustrate this functionality.

Procedure

1. Navigate to the Fields page on the dataset page.
2. Click Edit Fields.

The screenshot shows the Cloudera Data Visualization interface. The top navigation bar includes 'HOME', 'VISUALS', and 'DATA'. The left sidebar contains a 'Dataset Detail' section with 'Related Dashboards' (1) and a 'Fields' tab (1). The main content area is titled 'Dataset: Restaurant Inspection SF' and 'Fields'. An 'EDIT FIELDS' button (2) is highlighted. Below this, there are two panels: 'Dimensions' and 'Measures'. The 'Dimensions' panel shows a list of fields under the 'restaurant_scores_lives_standard' dataset, including business_id, business_name, business_address, business_city, business_state, business_postal_code, business_latitude, business_longitude, business_location, business_phone_number, inspection_id, inspection_date, inspection_type, violation_id, violation_description, and risk_category. The 'Measures' panel shows a list of fields under the 'restaurant_scores_lives_standard' dataset, including inspection_score.

3. Click Edit Field for the field that you want to use as an aggregate.



4. Choose a Default Aggregation type from the drop-down list.

Edit Field Parameters

Basic Settings

Display Format

Color

Base Column: inspection_score

Display Name

inspection_score

Field Comment

Enter field comment

Default Aggregation

Average

Geo Type

None

☒ Show field in data detail screen

☒ Show field in Visual Designer

☐ Use as a partition column for Analytical Views

Category

☐ Dimension ☒ Measure

CANCEL

APPLY

5. Click APPLY.
6. Navigate to the Search Modeling page.

7. Select the Default Aggregation Field from the drop-down list.

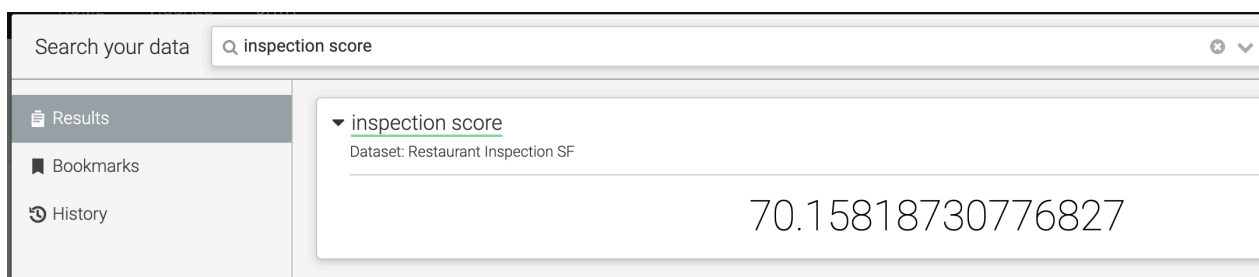
Default Aggregation Field

Specify the default aggregation field to display if none is explicitly identified in the search phrase.

8. Click SAVE.

Results

If you type 'inspection score' in the search bar you get an average score as a result.



Specifying suggested questions for a dataset in NLS

You can help users explore the dataset by pre-populating the dropdown with some search words and phrases in the dataset. You can also preview the results of the searches.

Procedure















1. Below Suggested Questions, click Add New Item.

2. In the text box, enter the search phrase.

In this example, several suggestions have been added.

Suggested Questions

List the questions the users see when using search for this dataset.

Suggestion	Try it	
top 10 violations		
trends by year		
top 10 restaurant violations by zip		
restaurant location where score is < 50		
restaurant trends in downtown sf as line		
top 10 restaurants where risk is "high risk"		
top 20 inspection count by business		

 Add New Item

3. Click the Search icon to preview the visual results.

A separate modal window appears, showing your results.

4. Click SAVE.



Note: To remove a search phrase, click the Trash icon next to it.

Specifying word substitutions for a dataset in NLS

Word substitutions enable you to connect common words with valid data specifications when using Natural Language Search (NLS) on Data Visualization datasets.

About this task

Common words and phrases have specific meaning in each dataset. In this example, you can see how to specify the meaning of the phrase 'current year' and the word 'domestic' into internal search terms.

Procedure

1. Under Word Substitutions, click Add New Item.
2. In the new row, add a common speech component in the Input word or phrase column.


3. In the same row, under Internal search term, specify the equivalent term in the dataset.

In this example, the phrase 'current year' was added with the clause 'year:2019', setting the current year to 2019.

Word Substitutions

List simple word or phrase substitutions that translate into internal search terms.

Examples: "domestic" → "country:'US'", "this year" → "year:2021", "location" → "latitude and longitude"

Input word or phrase	Internal search term	
current year	year:2019	

[+ Add New Item](#)

4. Click Add New Item again, if you want to add more word substitutions.

Configuring date and time for search requirements

Before configuring time on the datasets in Data Visualization, you must navigate to the relevant datasets.

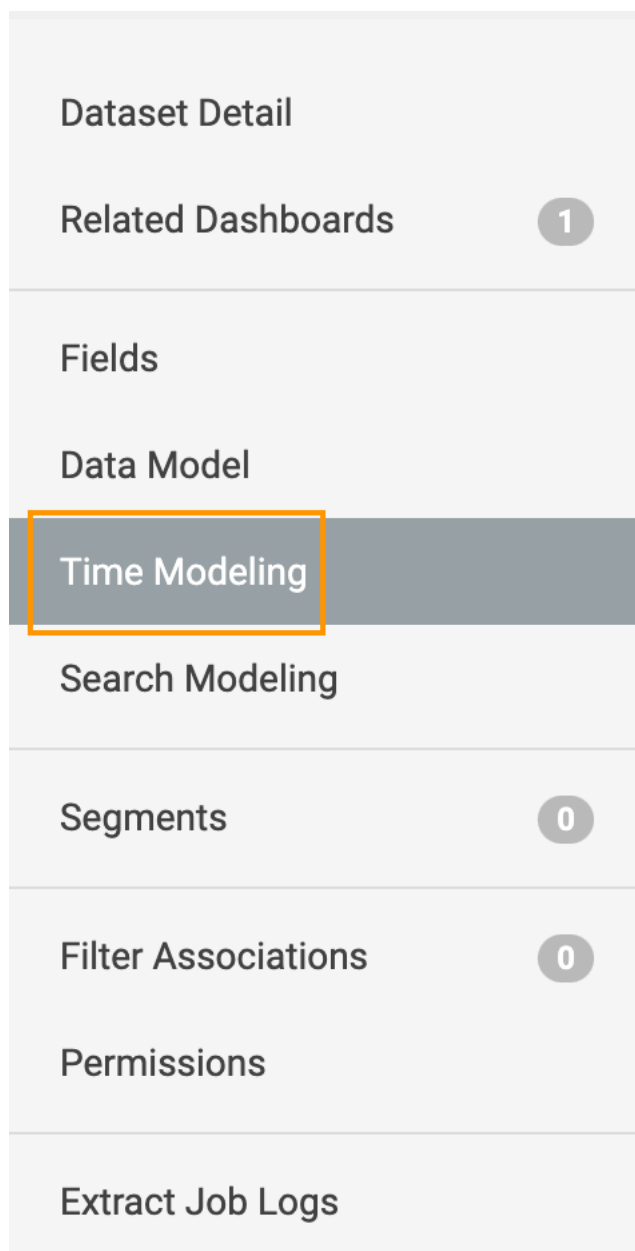
Procedure

1. On the main navigation bar, click DATA.



2. In the list of datasets, select the one you want to enable for NLS.

In this example, the Restaurant Inspections in SF dataset has been selected. The Dataset Detail page appears.



3. On the left navigation menu of the Dataset Detail page, click Time Modeling.

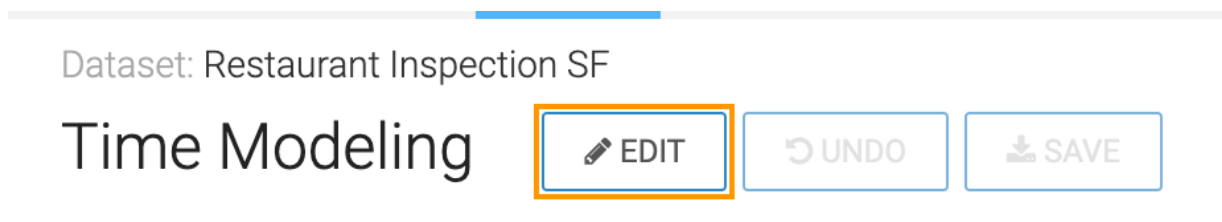
Specifying the date/time field and the format of date/time for dataset

About this task

In the Time Modeling page, specify the Date/Time Field of the dataset using the following steps:

Procedure

1. At the top of the page, click EDIT.



2. Under Date/Time Field, click the first drop-down list and select the name of the correct field, one that contains the date information.

The system groups dataset fields in timestamp or other time formats at the top, under the heading Date Fields.

In this example, the 'inspection_date' field has been selected.

Date/Time Field

Select the date/time field of the dataset, and specify its format. Dashboard time controls use this field.

A screenshot of the 'Date/Time Field' configuration section. It contains two drop-down menus. The first drop-down menu is set to 'inspection_date' and is highlighted with an orange rectangular box. The second drop-down menu is set to 'YYYY-MM-DD'.

3. Under Date/Time Column, click the second drop-down list and select the appropriate date format mask.

Date/Time Field

Select the date/time field of the dataset, and specify its format. Dashboard time controls use this field.

A screenshot of the 'Date/Time Field' configuration section. It contains two drop-down menus. The first drop-down menu is set to 'inspection_date'. The second drop-down menu is set to 'YYYY-MM-DD' and is highlighted with an orange rectangular box.

Depending on your location, the masks may have a different format. For example, the two options in this example are YYYY-MM-DD and YYYY-MM-DD HH:mm:ss. In this example, the dataset does not contain the hour, minute, and second information, so the default format YYYY-MM-DD is kept.



Note: If the date in the dataset is not in this format, a computed dataset column is needed that parses the date information into the desired format.

4. Click SAVE.

Specifying the default time interval in dataset

About this task

In the Time Modeling page, specify the default interval by following these steps:

Procedure

1. Under Date/Time Field, click the third drop-down list for Default Interval.

2. From the menu, select the appropriate interval. Depending on your location, you may have different options. For example, the three options in this example are Month, Quarter, and Year.

In the figure below Month is selected.

In natural language search, select the default time interval for analyzing metric trends

A rectangular dropdown menu with a light gray border. The word "Month" is displayed in a dark gray font on the left side. On the right side, there is a small, dark gray downward-pointing chevron icon.

3. Click SAVE.