



# BigML Cluster Cheat Sheet

## Sampling (Cont.)

## Cluster configuration

### Cluster configuration options

Option	Description	Default	API Name
<b>K-means: number of clusters</b>	Specifies the number of clusters to be found. Choose from 2 up to 300 clusters.	8	k
<b>G-means: critical value</b>	Determines how strictly the data distribution in the neighborhood of a candidate cluster should fit a Gaussian distribution. The strictest fit corresponds to a value of 1. Higher critical values will tend to find fewer clusters. You can set a maximum value of 20.	5	critical_value
<b>Default numeric value</b>	Replaces missing numeric values in the dataset with the field maximum, mean, median, minimum, or zero. If you don't enable this option, your instances with missing numeric values will be ignored. However, if all the instances contain at least a missing value, BigML automatically replaces them with the median.	Null	default_numeric_value
<b>Scale fields</b>	Assigns scales to particular fields.	False	field_scales
<b>Auto-scaled</b>	Scales all numeric fields so their standard deviations are 1. This makes each field have roughly equivalent influence.	True	balance_fields
<b>Weights</b>	Assigns different weights to the instances in your dataset. Any numeric field with no negative or missing values is valid as a weight assignable field. Each instance will be weighted individually according to the weight field's value.	False	weight_field
<b>Summary fields</b>	Sets summary fields, which aren't used to compute clusters, but their values are included when you create a dataset from a cluster. Non-preferred fields aren't eligible as summary fields. If you want to include a non-preferred field as a summary field, you will first need set that field as preferred.	[]	summary_fields

### Sampling

Option	Description	Default	API Name
<b>Rate</b>	Allows you to set the proportion of the dataset you want to consider between 0% and 100%.	100%	sample_rate
<b>Range</b>	Specifies a linear subset of the dataset instances that you want to be considered for the sample (example: from instance 5 to instance 1,000). The rate you set will be computed over the range configured.	(1, max. rows in dataset)	range

## Output File Options

Option	Description	Default	API Name
<b>Fields separator</b>	Allows you to choose the best separator for your fields.	Comma	separator
<b>Show/hide fields</b>	Allows you to show or hide the rest of the fields in your output file.	True	output_fields
<b>Headers</b>	Allows you to show or hide the names of your columns in the output file.	True	header
<b>Centroid column name</b>	Allows you to set the name for the centroid column in your output file.	Centroid	centroid_name
<b>Distance</b>	Allows you to include the distance for each centroid per instance.	False	distance
<b>Distance column name</b>	Allows you to set the name for the distance column in your output file.	Centroid	distance_name
<b>New line</b>	Sets the character to use as the line break in the generated csv file. "LF", "CRLF".	LF	newline

## Output Dataset

Option	Description	Default	API Name
<b>Output dataset</b>	Defines whether a dataset with the results should be automatically created or not.	True	output_dataset



## Batch Centroid Configuration

### Default Numeric Values

Option	Description	Default	API Name
<b>Default numeric value</b>	Replaces missing numeric values in your dataset by the field's maximum, mean, median, minimum, or zero. If you do not activate this option, your instances with missing numeric values will be ignored and you will not get a prediction for them.	Null	default_numeric_value