Paxata®

Adding Custom Cluster and Edit Functions

PAXATA RELEASE 2.19 | DOCUMENT REV. 1

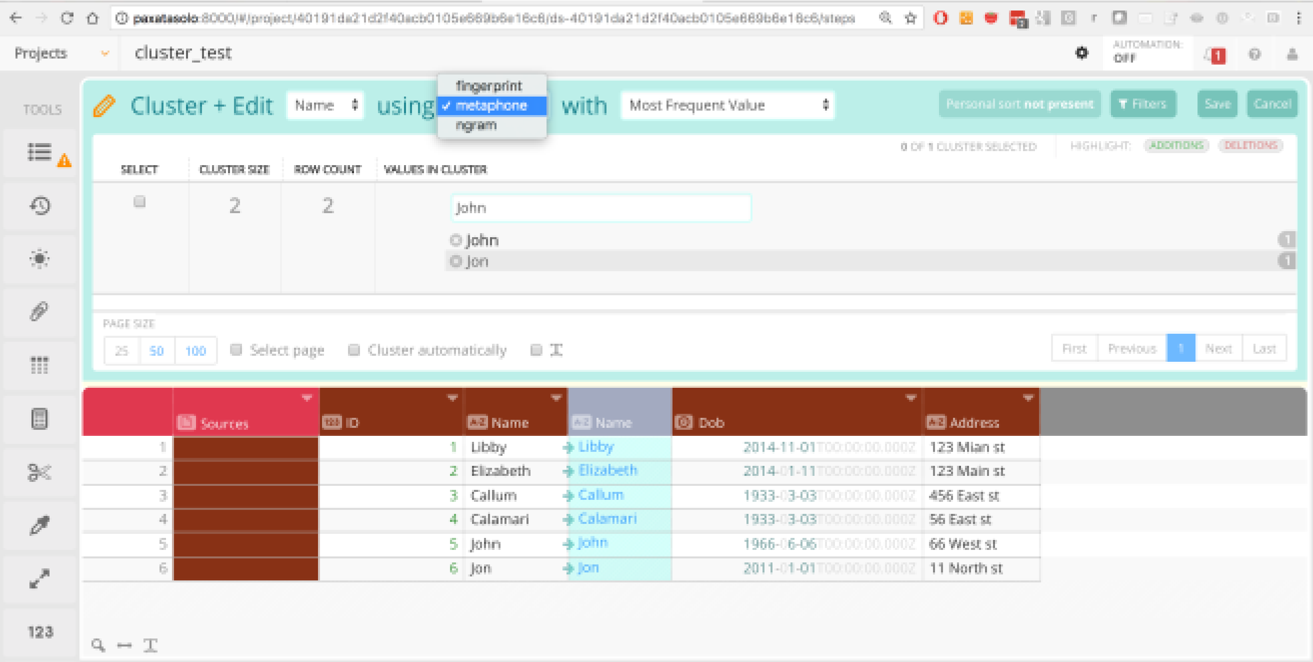
INTRODUCTION

Occasionally custom Cluster + Edit functions are required to achieve the desired clustering results (typically when there are very specific requirements). This document outlines the high-level steps to integrate a new function.

Screenshots below.

Questions – callum@paxata.com

When a *Project* is created, one of the functions I can perform is a **Cluster + Edit**, as the screenshot shows. Out of the box, I have 3 strategies I can use to perform this (**fingerprint, metaphone** and **ngram**):



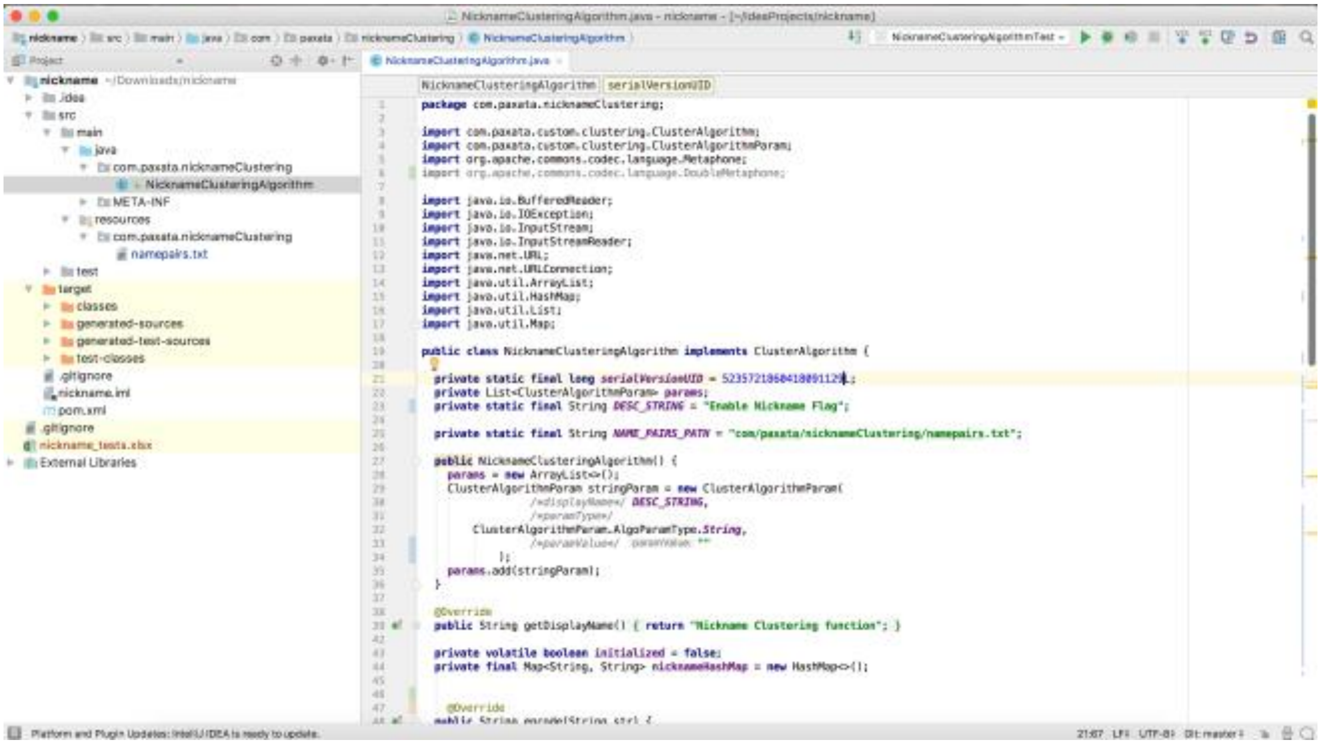
Cluster + Edit Name using metaphone with Most Frequent Value

SELECT	CLUSTER SIZE	ROW COUNT	VALUES IN CLUSTER
□	2	2	John ○ John 1 ○ Jon 1

□ Select page □ Cluster automatically □

	Sources	ID	Name	Name	Dob	Address
1		1	Libby	Libby	2014-11-01T00:00:00.000Z	123 Mian st
2		2	Elizabeth	Elizabeth	2014-01-11T00:00:00.000Z	123 Main st
3		3	Callum	Callum	1933-03-03T00:00:00.000Z	456 East st
4		4	Calamari	Calamari	1933-03-03T00:00:00.000Z	56 East st
5		5	John	John	1966-06-06T00:00:00.000Z	66 West st
6		6	Jon	Jon	2011-01-01T00:00:00.000Z	11 North st

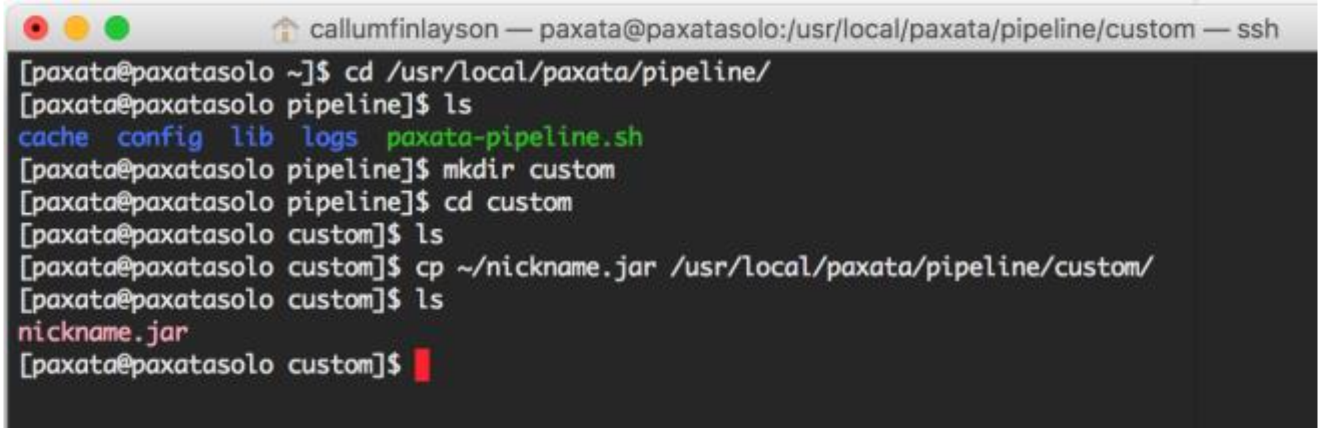
You can use a different strategy, by including a custom **Cluster + Edit** function. This is done by writing java code to create a custom JAR (Java ARchive) file. (Paxata provides the sample code on how to do this). Below is an example of some custom code:


NicknameClusteringAlgorithm.java - nickname - [~/IdeaProjects/nickname]
nickname
src
main
java
com
paxata
nicknameClustering
NicknameClusteringAlgorithm
NicknameClusteringAlgorithmTest
Project
nickname ~/Downloads/nickname
.idea
src
main
java
com.paxata.nicknameClustering
NicknameClusteringAlgorithm
META-INF
resources
com.paxata.nicknameClustering
namepairs.txt
test
target
classes
generated-sources
generated-test-sources
test-classes
.gitignore
nickname.iml
pom.xml
.gitignore
nickname_tests.xlsx
External Libraries
NicknameClusteringAlgorithm.java
NicknameClusteringAlgorithm
serialVersionUID
Platform and Plugin Updates: IntelliJ IDEA is ready to update.
21:67 LF UTF-8 Git: master

```java
package com.paxata.nicknameClustering;

import com.paxata.custom.clustering.ClusterAlgorithm;
import com.paxata.custom.clustering.ClusterAlgorithmParam;
import org.apache.commons.codec.language.Metaphone;
import org.apache.commons.codec.language.DoubleMetaphone;

import java.io.BufferedReader;
import java.io.IOException;
import java.io.InputStream;
import java.io.InputStreamReader;
import java.net.URL;
import java.net.URLConnection;
import java.util.ArrayList;
import java.util.HashMap;
import java.util.List;
import java.util.Map;

public class NicknameClusteringAlgorithm implements ClusterAlgorithm {

  private static final long serialVersionUID = 5235721868418091129L;
  private List<ClusterAlgorithmParam> params;
  private static final String DESC_STRING = "Enable Nickname Flag";

  private static final String NAME_PAIRS_PATH = "com/paxata/nicknameClustering/namepairs.txt";

  public NicknameClusteringAlgorithm() {
    params = new ArrayList<>();
    ClusterAlgorithmParam stringParam = new ClusterAlgorithmParam(
            /*displayName*/ DESC_STRING,
            /*paramType*/
        ClusterAlgorithmParam.AlgoParamType.String,
            /*paramValue*/ paramValue: ""
          );
    params.add(stringParam);
  }

  @Override
  public String getDisplayName() { return "Nickname Clustering function"; }

  private volatile boolean initialized = false;
  private final Map<String, String> nicknameHashMap = new HashMap<>();


    @Override
  public String encode(String str) {
```

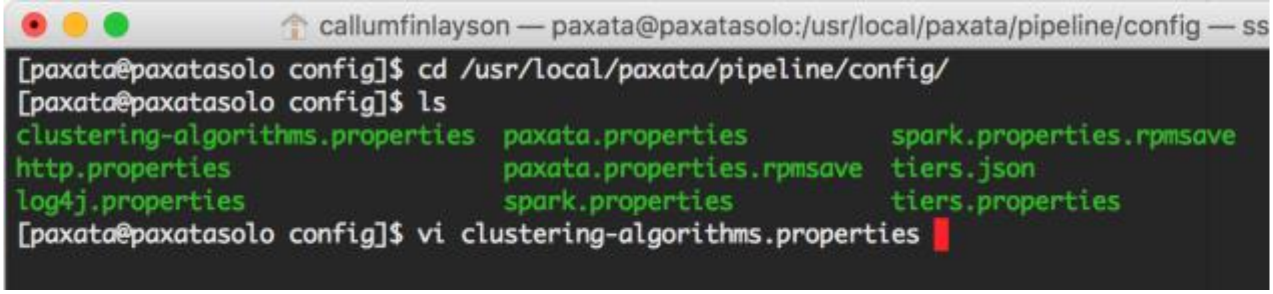
Once this is compiled and a JAR file has been created we need to perform 3 things to make this active:

1) Copy the JAR file (in this case “nickname.jar” to the correct location on the pipeline server). It is recommended that we create a new directory called “custom” for our custom cluster + edit JAR files:


callumfinlayson — paxata@paxatasolo:/usr/local/paxata/pipeline/custom — ssh

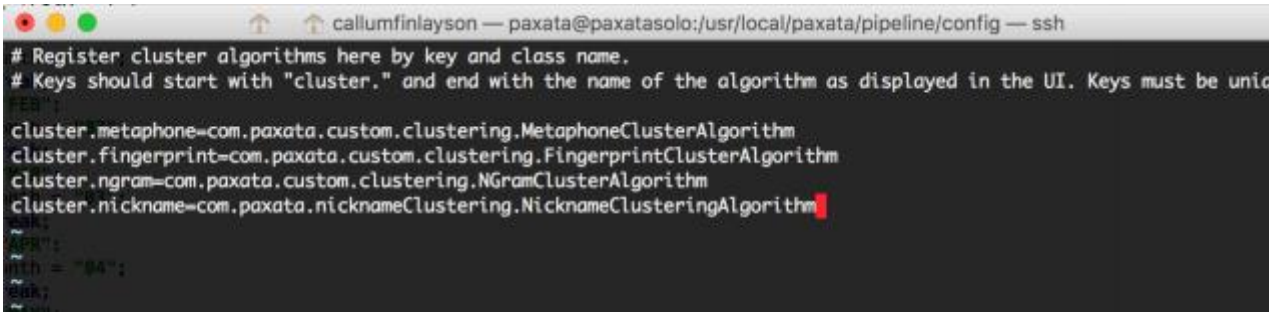
```
[paxata@paxatasolo ~]$ cd /usr/local/paxata/pipeline/
[paxata@paxatasolo pipeline]$ ls
cache  config  lib  logs  paxata-pipeline.sh
[paxata@paxatasolo pipeline]$ mkdir custom
[paxata@paxatasolo pipeline]$ cd custom
[paxata@paxatasolo custom]$ ls
[paxata@paxatasolo custom]$ cp ~/nickname.jar /usr/local/paxata/pipeline/custom/
[paxata@paxatasolo custom]$ ls
nickname.jar
[paxata@paxatasolo custom]$ 
```

2) Navigate the ***clustering-algorithms.properties*** file to tell Paxata to add a new custom **cluster + edit** function:


callumfinlayson — paxata@paxatasolo:/usr/local/paxata/pipeline/config — ss

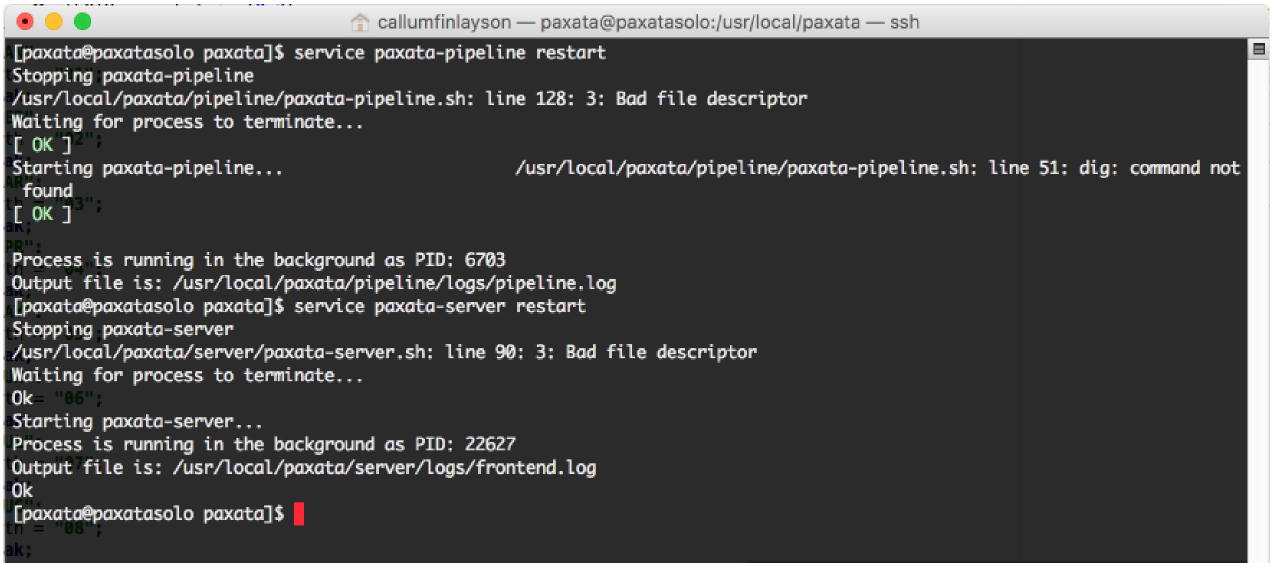
```
[paxata@paxatasolo config]$ cd /usr/local/paxata/pipeline/config/
[paxata@paxatasolo config]$ ls
clustering-algorithms.properties  paxata.properties          spark.properties.rpmsave
http.properties                   paxata.properties.rpmsave  tiers.json
log4j.properties                  spark.properties           tiers.properties
[paxata@paxatasolo config]$ vi clustering-algorithms.properties
```

Then add a single line in the ***clustering-algorithms.properties*** file (this will depend on how the JAR file has been packaged). The text I am using is *"cluster.nickname=com.paxata.nicknameClustering.NicknameClusteringAlgorithm"* (as displayed below):


callumfinlayson — paxata@paxatasolo:/usr/local/paxata/pipeline/config — ssh

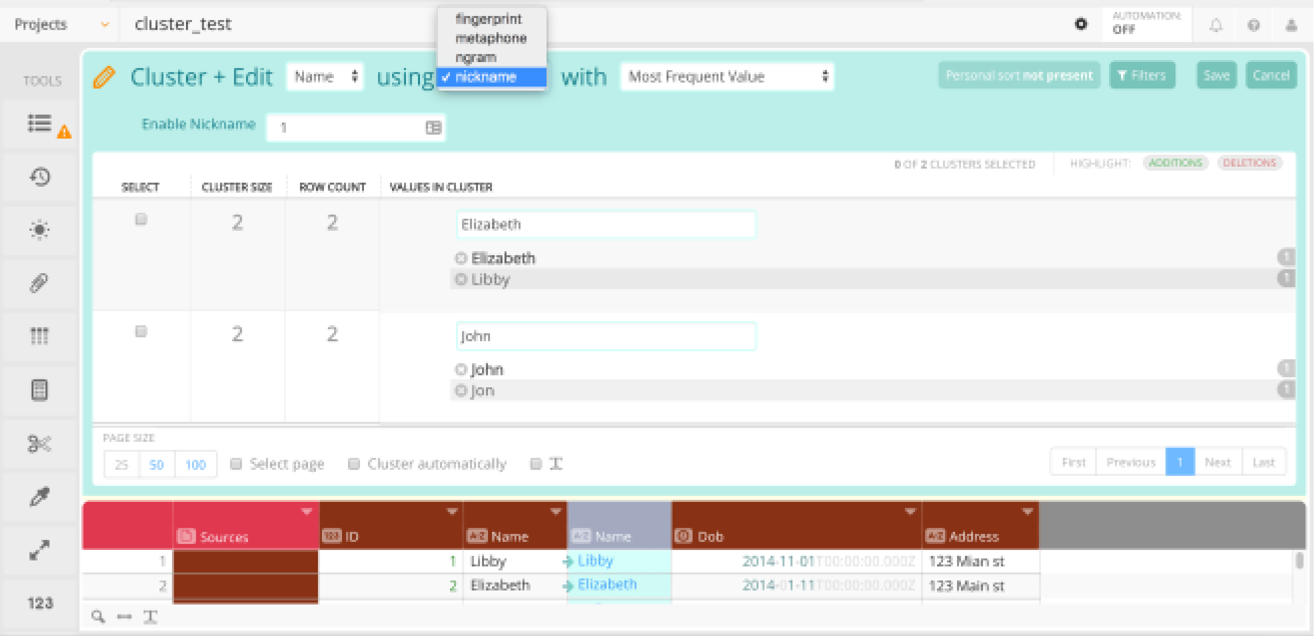
```
# Register cluster algorithms here by key and class name.
# Keys should start with "cluster." and end with the name of the algorithm as displayed in the UI. Keys must be uniq

cluster.metaphone=com.paxata.custom.clustering.MetaphoneClusterAlgorithm
cluster.fingerprint=com.paxata.custom.clustering.FingerprintClusterAlgorithm
cluster.ngram=com.paxata.custom.clustering.NGramClusterAlgorithm
cluster.nickname=com.paxata.nicknameClustering.NicknameClusteringAlgorithm
~
~
~
~
```

3) Finally, restart the pipeline and the core server by typing *"service paxata-pipeline restart",* followed by *"service paxata-server restart":*



```
[paxata@paxatasolo paxata]$ service paxata-pipeline restart
Stopping paxata-pipeline
/usr/local/paxata/pipeline/paxata-pipeline.sh: line 128: 3: Bad file descriptor
Waiting for process to terminate...
[ OK ]
Starting paxata-pipeline...                    /usr/local/paxata/pipeline/paxata-pipeline.sh: line 51: dig: command not
 found
[ OK ]

Process is running in the background as PID: 6703
Output file is: /usr/local/paxata/pipeline/logs/pipeline.log
[paxata@paxatasolo paxata]$ service paxata-server restart
Stopping paxata-server
/usr/local/paxata/server/paxata-server.sh: line 90: 3: Bad file descriptor
Waiting for process to terminate...
Ok
Starting paxata-server...
Process is running in the background as PID: 22627
Output file is: /usr/local/paxata/server/logs/frontend.log
Ok
[paxata@paxatasolo paxata]$
```

This will add a new **Cluster + Edit** algorithm to use in your Paxata installation (in this case called *“nickname”*):



Cluster + Edit Name using nickname with Most Frequent Value

Enable Nickname 1

0 OF 2 CLUSTERS SELECTED

SELECT	CLUSTER SIZE	ROW COUNT	VALUES IN CLUSTER
□	2	2	Elizabeth ○ Elizabeth 1 ○ Libby 1
□	2	2	John ○ John 1 ○ Jon 1

PAGE SIZE 25 50 100 □ Select page □ Cluster automatically □

	Sources	ID	Name	Name	Dob	Address
1		1	Libby	Libby	2014-11-01T00:00:00.000Z	123 Mian st
2		2	Elizabeth	Elizabeth	2014-01-11T00:00:00.000Z	123 Main st

Companies around the globe rely on Paxata to get smart about information. Paxata is the pioneer that intelligently empowers all business consumers to transform raw data into ready information, instantly and automatically, with an enterprise-grade, self-service data preparation application and machine learning platform. Our Adaptive Information Platform weaves data into an information fabric from any source and any cloud to create trusted insights. Business consumers use clicks, not code to achieve results in minutes, not months. With Paxata, Be an Information Inspired Business.

Paxata is headquartered in Redwood City, California with offices in New York, Ohio, Washington D.C., and Singapore.

Paxata®

Paxata Headquarters 305 Walnut Street Redwood City, CA 94063 1-855-9-PAXATA paxata.com

in You Tube f