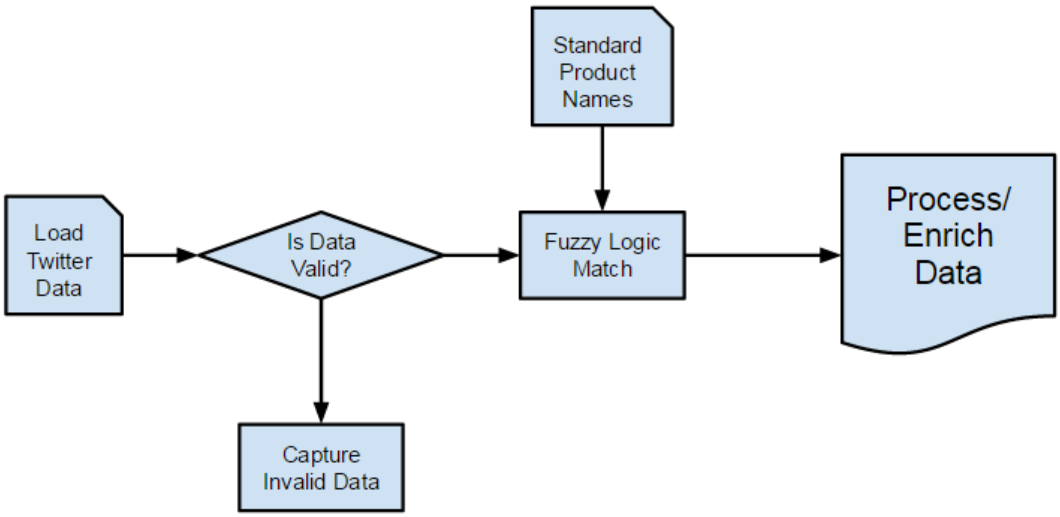
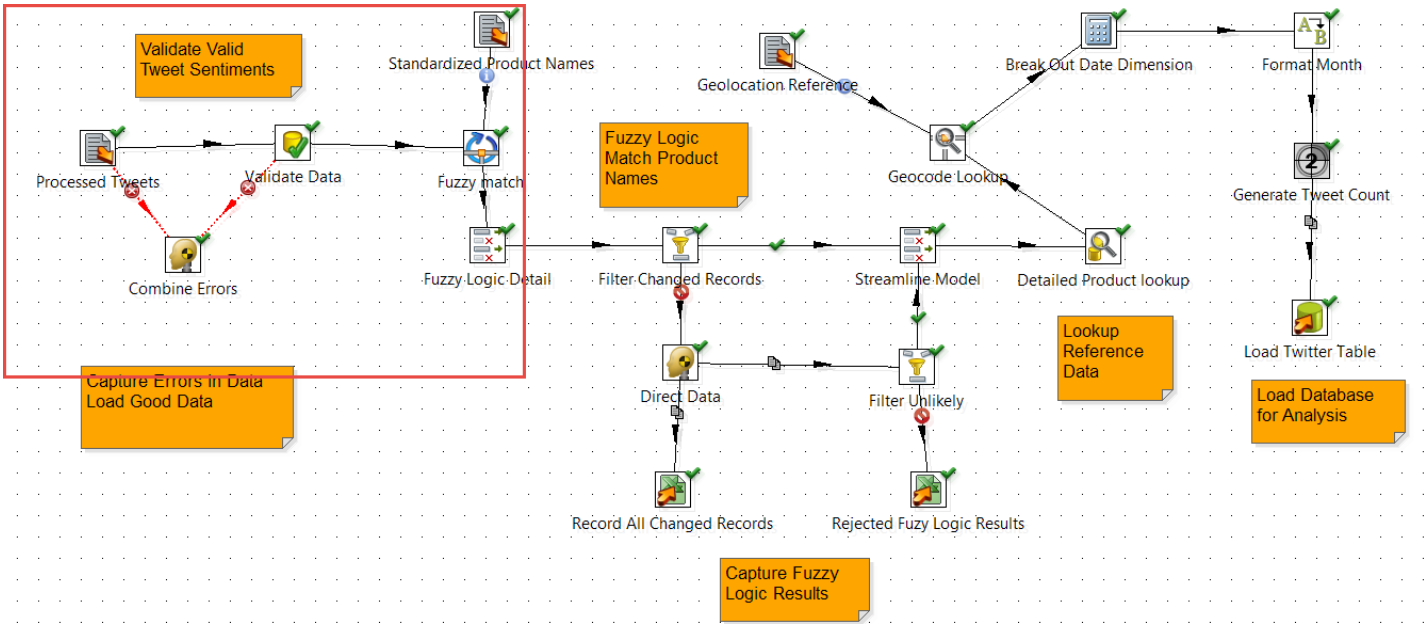
Data Cleansing and Validation

In this exercise, a mobile phone company is working with sentiment data from tweets about their products. They would like to ingest the data, validate whether it against some custom rules, and use fuzzy logic matching to match what products people are tweeting about to our list of standardized product names.



The workflow above, once implemented in Pentaho Data Integration, will result in a transformation that looks like the following:



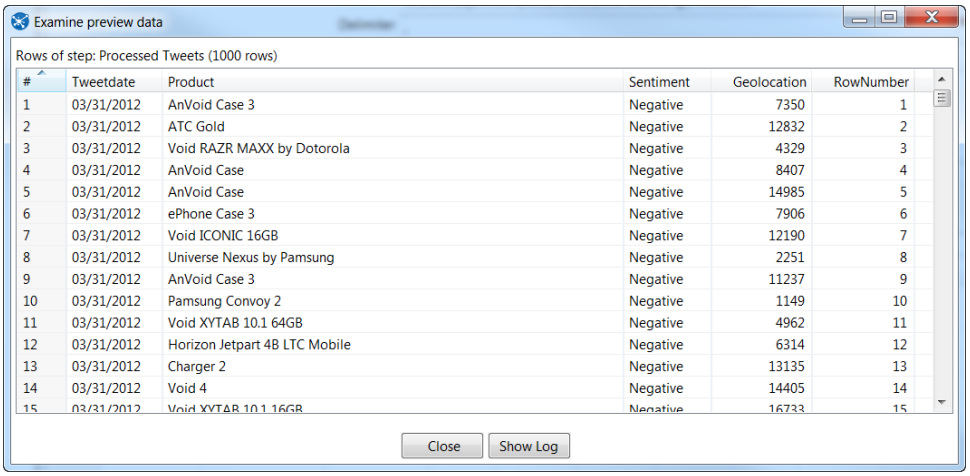
Build the Data Cleansing and Validation Exercise

1. Open the “data_cleansing_and_validation_final” transformation from the Desktop/WorkshopTraining/05_data_cleansing_and_validation folder for reference purposes.

2. Open the “data_cleansing_and_validation_start” transformation from the Desktop/WorkshopTraining/05_data_cleansing_and_validation folder.

3. Save “data_cleansing_and_validation_start” as “data_cleansing_and_validation_student” into the /Desktop/WorkshopTraining/student_files/05_data_cleansing_and_validation folder by selecting File -> Save As.

4. Open the Processed Tweets step and preview the data

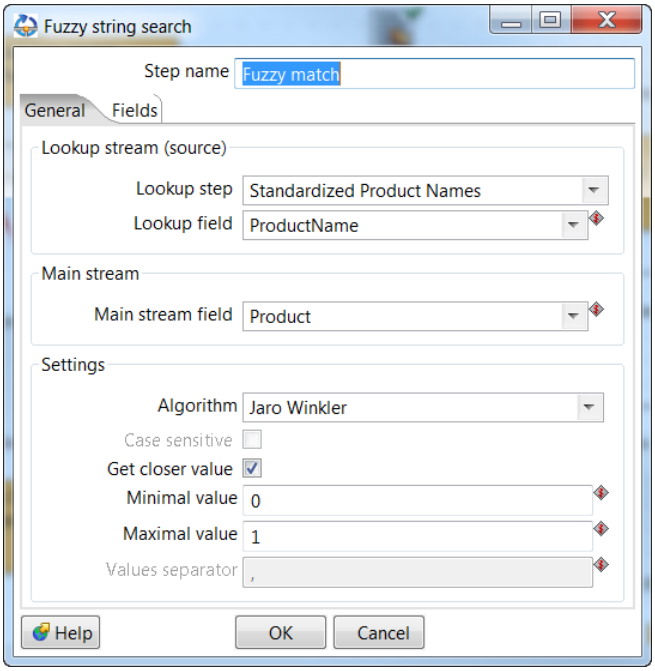


#	Tweetdate	Product	Sentiment	Geolocation	RowNumber
1	03/31/2012	AnVoid Case 3	Negative	7350	1
2	03/31/2012	ATC Gold	Negative	12832	2
3	03/31/2012	Void RAZR MAXX by Dotorola	Negative	4329	3
4	03/31/2012	AnVoid Case	Negative	8407	4
5	03/31/2012	AnVoid Case	Negative	14985	5
6	03/31/2012	ePhone Case 3	Negative	7906	6
7	03/31/2012	Void ICONIC 16GB	Negative	12190	7
8	03/31/2012	Universe Nexus by Pamsung	Negative	2251	8
9	03/31/2012	AnVoid Case 3	Negative	11237	9
10	03/31/2012	Pamsung Convoy 2	Negative	1149	10
11	03/31/2012	Void XYTAB 10.1 64GB	Negative	4962	11
12	03/31/2012	Horizon Jetpart 4B LTC Mobile	Negative	6314	12
13	03/31/2012	Charger 2	Negative	13135	13
14	03/31/2012	Void 4	Negative	14405	14
15	03/31/2012	Void XYTAB 10.1 16GB	Negative	16733	15

5. Open the Validate Data step click New Validation. Call it Sentiment Validation and enter the following values:

Field	Value
Report all errors, not only the first	Select checkbox
Validation Description	Sentiment Validation
Name of Field to Validate	Sentiment
Error code	DataValidator001
Error Description	Not Found in Expected Set
Verify data type?	Select checkbox
Data type	String
Null Allowed?	Uncheck this box
Allowed values	Positive Negative Neutral

6. Open the Fuzzy match step and enter the following values:



7. Save your transformation and run it…you will notice the following errors

```
2015/03/04 13:07:02 - Validate Data.0 - ERROR (version 5.2.0.0, build 1 from 2014-09-30_19-48-28 by buildguy) : Unexpected error
2015/03/04 13:07:02 - Validate Data.0 - ERROR (version 5.2.0.0, build 1 from 2014-09-30_19-48-28 by buildguy) : org.pentaho.di.core.exception.KettleException:
2015/03/04 13:07:02 - Validate Data.0 - Not Found in Expected Set
2015/03/04 13:07:02 - Validate Data.0 - Not Found in Expected Set
```

Notice that the error states “Not Found in Expected Set”. These errors will cause the transformation to stop completely. Now we are going to implement error handling so that we can capture these errors and continue processing data.

Error Handling

1. In the Design Tab search box, type in dummy and you will see the Dummy step listed below, drag it underneath the Validate Data step

2. Connect the Validate Data step to the Dummy (do nothing) step and select Error handling of step, then select the copy button.

3. Save and run the transformation again. You will notice another error:

```
2015/03/04 13:46:00 - Processed Tweets.0 - There were 1 conversion errors on line 9946
2015/03/04 13:46:00 - Processed Tweets.0 -
2015/03/04 13:46:00 - Processed Tweets.0 -
2015/03/04 13:46:00 - Processed Tweets.0 - Unexpected conversion error while converting value [Geolocation String] to an Integer
2015/03/04 13:46:00 - Processed Tweets.0 -
2015/03/04 13:46:00 - Processed Tweets.0 - Geolocation String : couldn't convert String to Integer
2015/03/04 13:46:00 - Processed Tweets.0 -
2015/03/04 13:46:00 - Processed Tweets.0 - Geolocation String : couldn't convert String to number : non-numeric character found at position 1 for value [test]
```

4. This is being caused by a data type issue on line 9946. It appears that we have a string character in a field that is expecting an integer. To capture this error, we want to repeat what we did in step 2 above.

5. Connect the Process Tweets step to the Dummy (do nothing) step and select Error handling of step, then select the copy button.

6. Now save and run the transformation.

7. Click on the Dummy (do nothing) step then select the Preview data tab beneath the transformation in the Execution Results tab.

Execution Results

Execution History | Logging | Step Metrics | Performance Graph | Metrics | Preview data

First rows | Last rows | Off

#	Tweetdate	Product	Sentiment	Geolocation	RowNumber	Error_Description	Error_FieldName	Error_Code
1	05/09/2012	Santech Phone 2	<null>	2646	5517	Not Found in Expected Set	Sentiment	DataValidator001
2	06/10/2012	ePad Dock 3	Neutral	<null>	9945	<null>	<null>	<null>
3	09/09/2012	ePhone Dock 3	Negative	<null>	22583	<null>	<null>	<null>
4	01/11/2013	Tasio 3	test	5542	39586	Not Found in Expected Set	Sentiment	DataValidator001

Review Exercise #2

What We Covered…

- Validating contextual data.
- Validating data types.
- Fuzzy logic matching.
- Capturing data
- Previewing data