

October 2008

Welcome to the October 2008 edition of IBM InfoSphere Information Server Developer's Notebook. This month we answer the question;

How do I make use of the QualityStage component to Information Server, more specifically, How do I create and use QualityStage custom rule sets?

Excellent question! While data cleansing and best record survivorship are the most commonly discussed and/or used functions within QualityStage, this component to Information Server does so much more. We regularly use QualityStage custom rule sets to parse free form text and output standardized data, allowing us to enrich data we previously owned.

QualityStage custom rule sets are easy to learn with just a small amount of practice. Using a phased approach, we will first discuss the Investigate stage of QualityStage in this edition of IISDN (and also overview QualityStage as a whole), then move on to the Standardize stage and custom rule sets in the next edition.

Software versions

All of these solutions were *developed and tested* on (IBM) InfoSphere Information Server (IIS) version 8.1, using the Microsoft Windows XP/SP2 platform to support IIS client programs, and a RedHat Linux Advanced Server 4 (RHEL 4) FixPak U6 32 bit SMP server (Linux kernel version 2.6.9-67.EL-smp) to support the IIS server components.

IBM InfoSphere Information Server allows for a single, consistent, and accurate view of data across the full width of the corporate enterprise, be it relational or non-relational, staged or live data. As a reminder, the IBM InfoSphere Information Server product contains the following major components;

WebSphere Business Glossary Anywhere™, WebSphere Information Analyzer™, WebSphere Information Services Director™, WebSphere DataStage™, WebSphere QualityStage™, WebSphere Metadata Server and Metabridges™, WebSphere Metadata Workbench™, WebSphere Federation Server™, Classic Federation™, Event Publisher™, Replication Server™, Rational Data Architect™, DataMirror Transformation Server™, and others.

Obviously, IBM InfoSphere Information Server is a large and capable product, addressing many strategic needs across the enterprise, and supporting different roles and responsibilities.

22.1 Terms and core concepts

The QualityStage component to IBM InfoSphere Information Server provides additional (high function) stages that appear within the integrated DataStage/QualityStage Designer program. These additional stages appear within the Data Quality drawer of the Palette view, and include;

- Investigate stage
- Standardize stage
- Match Frequency stage
- Unduplicate stage
- Reference Match stage
- Survive stage

Use of the Investigate stage, part of the QualityStage component of Information Server, is entirely optional. On some level, the Investigate stage could be viewed as a poor man's version of Information Analyzer, another component of Information Server. Further comments include;

- The Information Analyzer component of Information Server is 'team based', and all of the data discovery information it produces is promoted to the shared Metadata Repository. Information Analyzer has numerous graphical screens to track the discovery of data; analyzing data types, completeness of data, input field masks, and more.

Similar data (the similar result set) produced by the Investigate stage is not promoted to the Metadata Repository.

- Information Analyzer has an automatic and ongoing data measurement (base line data) feature towards input/received data, whereas the Investigate stage would have to have Jobs manually created to monitor same.
- However, the Investigate stage works with QualityStage custom rule sets and is handy when producing those; helping one understand how far you've gotten into the rules generation of the input data format and related.
- Net/net, Information Analyzer is the data discovery and monitoring component to Information Server, and QualityStage, including the Investigate stage, is the data standardization and cleansing component.

Use of the Standardize stage and by implication, Standardization Rules, are the gateway to the whole of the QualityStage component of Information Server. In effect, the Standardize stage takes a single or set of input columns, parses them, and minimally outputs *bucketed data*; meaning, the Standardize stage will take

free form text of a street address and say, 'this is the house number, this is the street name, this is the street directional', and so on.

There are pre-built Standardization Rules that come with QualityStage, and then you can build your own custom rule sets. The built in QualityStage Standardization Rules, and use thereof, is pretty well outlined within the Information Server documentation set. Building your own custom rules sets is the topic of the next edition of this document, (IBM) InfoSphere Information Server Developer's Notebook (IISDN) November/2008.

Figure22-1 displays the sample data set we will refer to in this, and the following month's edition of IISDN.

PJ01_InvestigateStage..InputFile.toCopy - Data Browser

CustomerNumber	FirstName	LastName	Company	Address1
101	Ludwig	Pauli	All Sports Supplies	213 Erstwild Court
102	Carole	Sadler	The Sports Spot	785 Geary St
103	Philip	Currie	Phil's Sports	654 Poplar
104	Anthony	Higgins	Play Ball!	East Shopping Cntr.
105	Raymond	Vector	Los Altos Sports	1899 La Loma Drive
106	George	Watson	Watson & Son	1143 Carver Place
107	Charles	Ream	Athletic Supplies	41 Jordan Avenue
108	Donald	Quinn	Quinn's Sports	587 Alvarado
109	Jane	Miller	Sport Stuff	Mayfair Mart
110	Roy	Jaeger	AA Athletics	520 Topaz Way
111	Frances	Keyes	The Sports Center	3199 Sterling Court
112	Margaret	Lawson	Runners & Others	234 Wyandotte Way
113	Lana	Beatty	Sportstown	654 Oak Grove
114	Frank	Albertson	The Sporting Place	947 Waverly Place
115	Alfred	Grant	Gold Medal Sports	776 Gary Avenue
116	Jean	Parmelee	Olympic City	1104 Spinosa Drive
117	Arnold	Sipes	Kids Korner	850 Lytton Court
118	Dick	Baxter	Blue Ribbon Sports	5427 College
119	Bob	Shorter	The Triathletes Club	2405 Kings Highway
120	Fred	Jewell	Century Pro Shop	6627 N. 17th Way
121	Jason	Wallack	City Sports	Lake Biltmore Mall
122	Cathy	O'Brian	The Sporting Life	543 Nassau Street
123	Marvin	Hanlon	Bay Sports	10100 Bay Meadows Ro
124	Chris	Putnum	Putnum's Putterers	4715 S.E. Adams Blvd
125	James	Henry	Total Fitness Sports	1450 Commonwealth Av
126	Eileen	Neelie	Neelie's Discount Sp	2539 South Utica Str
127	Kim	Satifer	A Big Blue Bike Shop	Blue Island Square
128	Frank	Lessor	Phoenix University	Athletic Department

Figure 22-1 Sample data set we refer to in this edition of IISDN, page 1 of 2.

And Figure 22-2 displays the remainder of this data.

Address2	City	State	ZipCode	Phone
	Sunnyvale	CA	94086	789-8075
	San Francisco	CA	94117	(415)822-1289
P. O. Box 3498	Palo Alto	CA	94303	(415)328-4543
422 Bay Road	Redwood City	CA	94026	(415)368-1100
	Los Altos	CA	94022	(415)776-3249
	Mountain View	CA	94063	(415)389-8789
	Palo Alto	CA	94304	(415)356-9876
	Redwood City	CA	94063	(415)544-8729
7345 Ross Blvd.	Sunnyvale	CA	94086	723-8789 ex 705
	Redwood City	CA	94062	(415)743-3611
	Sunnyvale	CA	94085	277-7245
	Los Altos	CA	94022	(415)887-7235 # 555
	Menlo Park	CA	94025	(415)356-9982
	Redwood City	CA	94062	(415)886-6677
	Menlo Park	CA	94025	
	Mountain View	CA	94040	(415)534-8822
	Redwood City	CA	94063	
	Oakland	CA	94609	(415)655-0011
	Cherry Hill	NJ	08002	609-663-6079
	Phoenix	AZ	85016	602-265-8754
350 W. 23rd Street	Wilmington	DE	19898	302-366-7511
	Princeton	NJ	08540	609-342-0054
Suite 1020	Jacksonville	FL	32256	904-823-4239
Suite 909C	Bartlesville	OK	74006	918-355-2074
	Brighton	MA	02135	617-232-4159
	Denver	CO	80219-40	303/936-7731 Extn 5
12222 Gregory Str	Blue Island	NY	60406	312-944-5691
1817 N. Thomas Rd	Phoenix	AZ	85008	602-533-1817

Figure 22-2 Sample data set we refer to in this edition of IISDN, page 2 of 2.

Comments related to the data set displayed in Figure22-1 and Figure22-2;

- Not wanting to have to type all of the data in ourselves, we grabbed the Customer (sample) table from our favorite database server, IBM Informix Dynamic Server.
- Really we will only refer to three of the columns above;
 - CompanyName-

A really good Company Name Standardization Rule (built in) comes with the QualityStage component to Information Server. Still, we'll use Company Name data to build our own custom rule.

This is the first custom rule we will build in the next edition of IISDN; our rule will strip common words from the leading portion of Company Names, for example,

The Sport's Spot

A Big Blue Bike Shop

become,

Sport's Spot, The

Big Blue Bike Shop, A

That's a handy rule to have for sorting your data sets.

In order to continue with the examples that follow, it would help if your sample data followed these existing data conditions.

- ZipCode-

The sample data in this column has the U.S. postal code in a 5 digit, and then also 5+4 digit format.

Again, a really good U.S. ZipCode (postal code) Standardization Rule comes with the QualityStage component of Information Server. We are merely using postal codes as a common example.

- Phone-

Here the sample data comes in a more complex set of formats, including,

(NULL)

(555)555-1212 Ex 555

(555)555-1212 555

(555)555-1212

555-1212 Ex 555

555-1212 555

555-1212

Or any combination of spaces and punctuation therein.

Your sample data, wherever and however you create it, should include each of the above examples.

Figure22-3 displays the sample DataStage/QualityStage Job we create for this example.

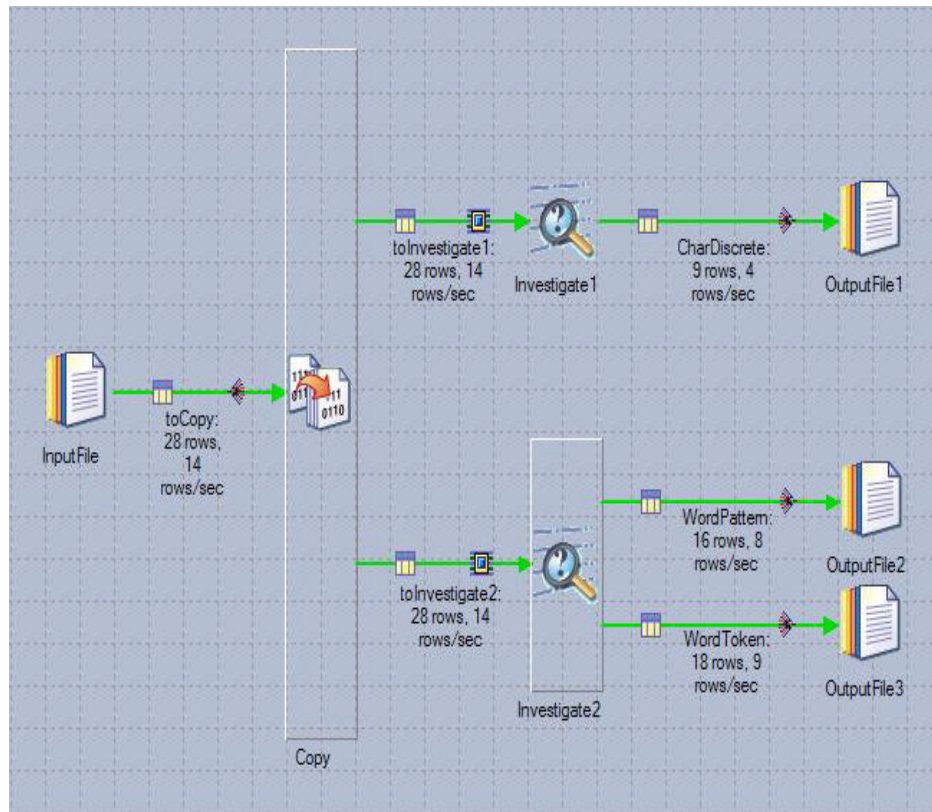


Figure 22-3 Sample DS/QS Job with Investigate stage.

Comments related to Figure22-3 include;

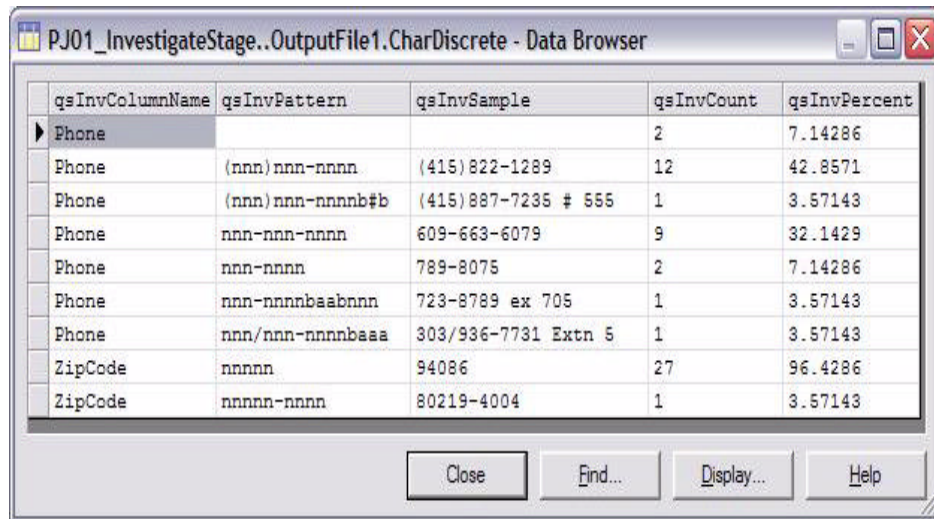
- The Investigate stage has 3 TABs, 3 types of primary reports that it can output; Character Discrete, Word Investigate, and Character Concatenate.

Today we are only going to discuss Character Discrete and Word Investigate, not Character Concatenate. (That's why you see two Investigate stages above, not three.)

We read from our source data file, and send this output to a Copy stage, allowing us to direct output to numerous (in this case, two) locations; two Investigate stages.

- The Word Investigate stage itself outputs two reports; Word Pattern, and Word Token; hence, this single stage offers two output locations.

Figure22-4 displays the result set from a Investigate stage, set for a Character Discrete report. (In this example, we called to Investigate only the Phone and ZipCode data from Figure22-2.)



qsInvColumnName	qsInvPattern	qsInvSample	qsInvCount	qsInvPercent
Phone			2	7.14286
Phone	(nnn)nnn-nnnn	(415)822-1289	12	42.8571
Phone	(nnn)nnn-nnnnb#b	(415)887-7235 # 555	1	3.57143
Phone	nnn-nnn-nnnn	609-663-6079	9	32.1429
Phone	nnn-nnnn	789-8075	2	7.14286
Phone	nnn-nnnnbaabnnn	723-8789 ex 705	1	3.57143
Phone	nnn/nnn-nnnnbaaa	303/936-7731 Extn 5	1	3.57143
ZipCode	nnnnn	94086	27	96.4286
ZipCode	nnnnn-nnnn	80219-4004	1	3.57143

Figure 22-4 Sample output from Investigate -> Character Discrete.

Comments related to Figure22-4 include;

- The result set displayed in Figure22-4 returns one line for every unique input data pattern it receives, per input column.

From the example above, ZipCode, in all of the data that was presented, has two unique patterns found to exist inside that data set,

94086 (or any numeric with just 5 spaces).

80219-4004 (or any 5 space numeric, followed by a hyphen, followed by a 4 space numeric).

For just ZipCode then, this tells us we have just two patterns we need to check for in any Standardization stage custom rule set we wish to create.

Phone offers a more complex input data set, with 7 patterns found to exist inside the input data set, including NULLs.

A good/complete custom rule set also handles any unknown, not previously anticipated patterns, to the input data set.

- The diagnostic columns returned from an Investigate stage, Character Discrete report include;
 - qsInvColumnName-

The name of the input column being reported on.

- **qsInvPattern-**

The observed input (masking) pattern. N, numeric, B blank, A alphanumeric, and then punctuation as displayed.

You can display patterns or literals here.

- **qsInvSample-**

A sample of the data being reported; one row is displayed by default, you may display multiples.

- **qsInvCount-**

The count of records in this grouping.

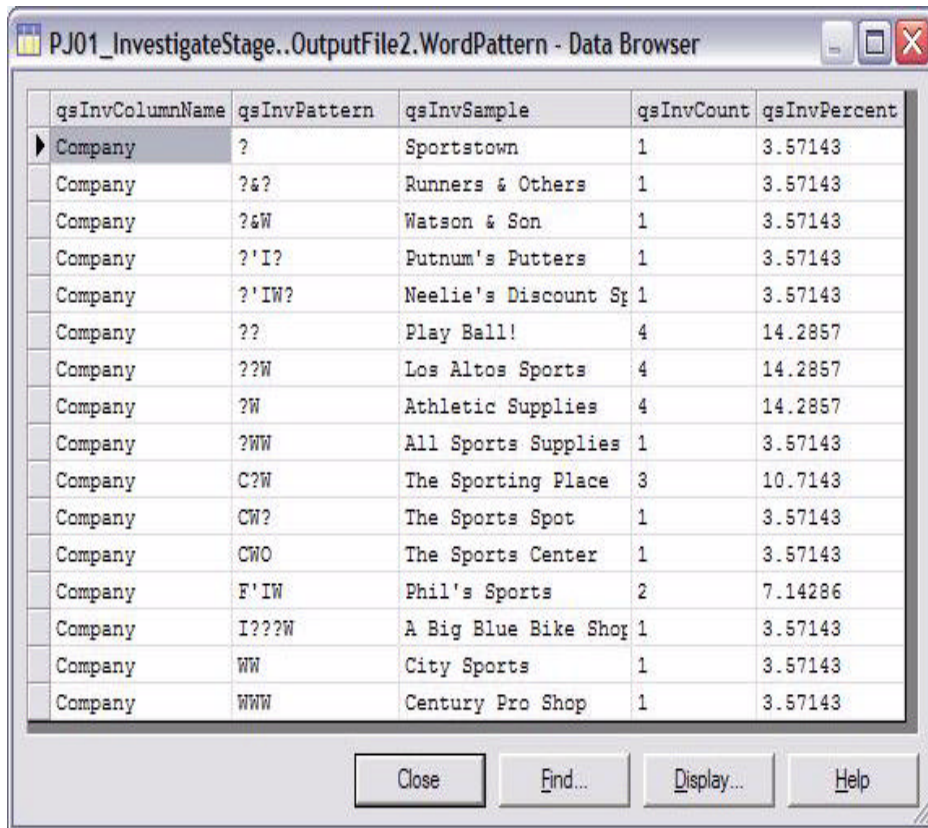
- **qsInvPercent-**

Count of records in this grouping over total records, a percentage.

Note: When the amount of data in this report becomes too cumbersome, you may of course send this output to a Sort stage, and/or Head and Tail stages, before writing to file. We tend to sort on `qsInvColumnName`, then `qsInvCount` descending.

There are also graphical reports available inside the Information Server Console program. That topic is not expanded upon further here.

Figure22-5 displays the result set from a Investigate stage, set for a Word Pattern report. (In this example, we called to Investigate only the Company (Name) data from Figure22-1.)



qsInvColumnName	qsInvPattern	qsInvSample	qsInvCount	qsInvPercent
Company	?	Sportstown	1	3.57143
Company	?&?	Runners & Others	1	3.57143
Company	?&W	Watson & Son	1	3.57143
Company	? 'I?	Putnum's Putters	1	3.57143
Company	? 'IW?	Neelie's Discount Sp	1	3.57143
Company	??	Play Ball!	4	14.2857
Company	??W	Los Altos Sports	4	14.2857
Company	?W	Athletic Supplies	4	14.2857
Company	?WW	All Sports Supplies	1	3.57143
Company	C?W	The Sporting Place	3	10.7143
Company	CW?	The Sports Spot	1	3.57143
Company	CWO	The Sports Center	1	3.57143
Company	F 'IW	Phil's Sports	2	7.14286
Company	I???W	A Big Blue Bike Shop	1	3.57143
Company	WW	City Sports	1	3.57143
Company	WWW	Century Pro Shop	1	3.57143

Figure 22-5 Sample output from Investigate -> Word Pattern.

Comments related to Figure22-5 include;

- The new column in Figure22-5 is the qsInvPattern column, aka, the Pattern Class column. *Pattern classes* include *simple classes* and *user-supplied classes*.

Simple classes include;

- ^ represents a wholly numeric token/word.
- + represents a wholly character token/word.
- > represents a mixed token/word, leading numeric.
- < represents a mixed token/word, leading character.

- @ represents a mixed token/word, numeric and character, either one leading, and anyone trailing. I.e., AA22GG. (This simple pattern class is used often for Canadian style postal codes.)
- & represents a single token/word of any type, best used when matched with another fixed token. I.e., give me any single next token/word following a known token/word, "Apt: 512B", etc.
- ? represents one or more unknown tokens/words. Similar in use to & above.
- ~ represents tokens/words leading with a punctuation character that is not used as a field separator. I.e., \$400.
- Generally all remaining punctuation is represented by its literal character; hyphens for hyphens, slash for slash, etc.

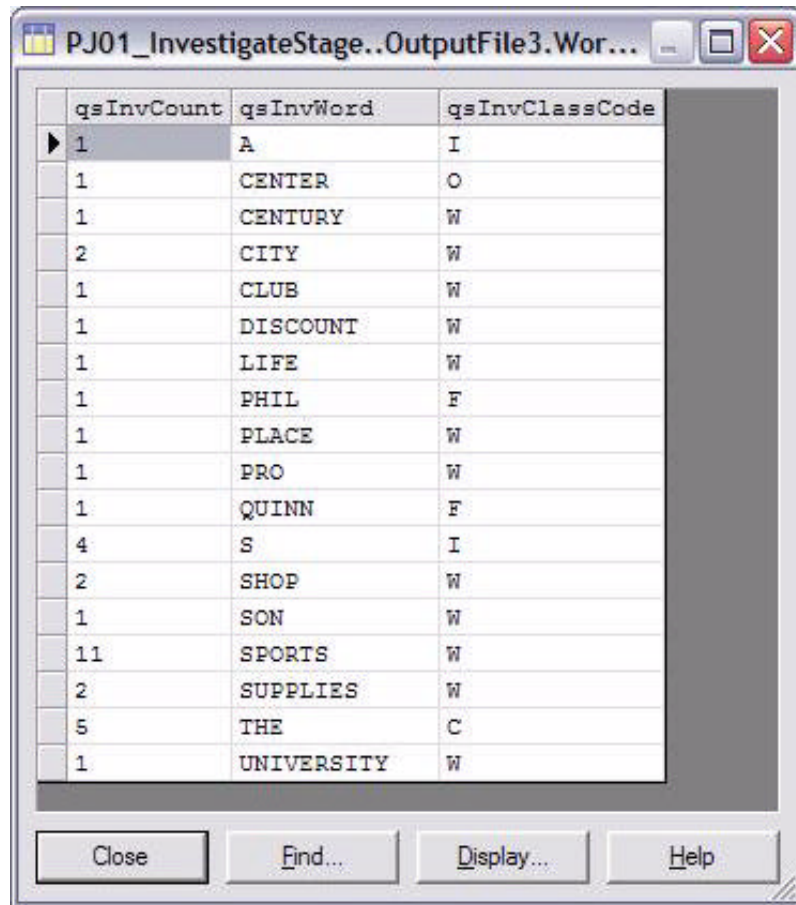
User-supplied classes are defined by the 26 alphabetic characters (A-Z). For example, one might create a user-supplied class for Directional (D), to include the words North, South,... N, S, E, W, NW, SE, etc. User Supplied classes can also include approximate or near spellings.

- In the qslnvPattern column, in Figure22-5 above, any alphabetic characters (A-Z) represent a user-supplied class that someone has defined.

This also implies that the Investigate -> Word Pattern report is associated with a QualityStage Rule Set, which it is; user-supplied classes being contained inside a rule set. To know what C, I, or W (from Figure22-5) represents, one has to check inside the applied rule set's Classification File.

For this example we used the QualityStage built in rule set for U.S. Company Names, detailed below.

Figure22-6 displays the result set from a Investigate stage, set for a Word Token report. (In this example, we called to Investigate only the Company (Name) data from Figure22-1.)



qsInvCount	qsInvWord	qsInvClassCode
1	A	I
1	CENTER	O
1	CENTURY	W
2	CITY	W
1	CLUB	W
1	DISCOUNT	W
1	LIFE	W
1	PHIL	F
1	PLACE	W
1	PRO	W
1	QUINN	F
4	S	I
2	SHOP	W
1	SON	W
11	SPORTS	W
2	SUPPLIES	W
5	THE	C
1	UNIVERSITY	W

Figure 22-6 Sample output from Investigate -> Word Token.

Comments related to Figure22-6 include;

- In Figure22-6 we see the individual words found to exist in a input data column, along with its associated token/word classification code.
- In Figure22-6 we might, for example, build a classification of common words like, “The”, “An”, etcetera, and use this report to gather our target values.

22.2 Creating the example outlined above

To create the three outputted reports detailed above, complete the following steps;

1. Using your favorite editor, create a sample data file equal to that as displayed in Figure22-1.

Minimally you should have 3 columns-

- Company (Name) should have entries that variably do and do not begin with a common word, for example,

The Sporting Spot (common word, The)

A Big Blue Bike Shop (common word, A)

Gold Medal Sports (no leading common word)

- ZipCode should include at least two patterns, for example,

53120

80129-4004

- Phone Number should present several patterns, including an extension, for example,

(303)470-0991 Ex 5

414-642-5555

642-5555

And more

2. Log on to the DataStage/QualityStage Designer program, create a new Parallel Job, and save that Job with a given name.

Drag and drop 7 (count) total stages from the Palette view onto the Parallel Canvas. Reposition and rename the stages as displayed in Figure22-3.

These 7 stages include;

- 4 (count) Sequential File stages.
- 2 (count) Investigate stages.
- 1 (count) Copy stage.

3. Set the properties on the input and output Sequential File stages. These include;

- Input/Output TAB -> Properties TAB -> Source -> File (name)
- Input/Output TAB -> Properties TAB -> Options -> First Line is Column Names -> True

(For output files only, or as appropriate.)

- Input/Output TAB -> Format TAB -> Record Level -> Final delimiter
- Input/Output TAB -> Format TAB -> Field defaults -> Delimiter
- Input/Output TAB -> Format TAB -> Field defaults -> Quote

- Input/Output TAB -> Columns

For input file only; the output files will automatically inherit their (output) columns.

Check to ensure that you can View Data on the input source file.

Note: The graphical reports one can generate from the output of the Investigate stage, require certain default properties on the output Sequential File stages; namely, comma delimited, column headers for the first line of data, etcetera.

As mentioned briefly above, these graphical reports are generated and managed inside the Information Server Console program. This topic is not expanded upon further here.

4. Configure the Copy stage-
 - Configure the Copy stage to output the Phone Number and ZipCode columns to the first Investigate stage, the one we will configure for a Character Discrete report.
 - Configure the Copy stage to output the Company (Name) to the second Investigate stage.
5. Configure the first Investigate stage, the one for a Character Discrete report-
 - a. Double-click the Investigate stage to open its Properties dialog box.

Click the TAB entitled, Character Discrete Investigate.

If the preceding Copy stage is configured correctly, two columns should be displayed as displayed in Figure22-7, (ZipCode and PhoneNumber).

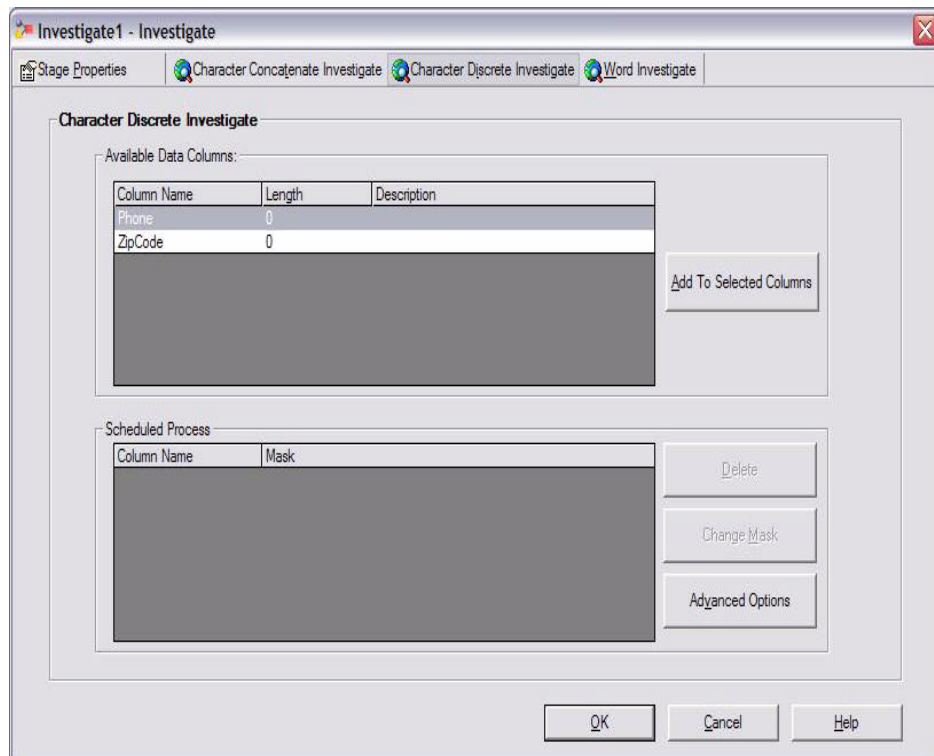


Figure 22-7 Investigate Stage -> Character Discrete Investigate TAB.

- b. Highlight the PhoneNumber or ZipCode columns, and Click the Add to Selected Columns button.

This action will produce the dialog box as displayed in Figure22-8.

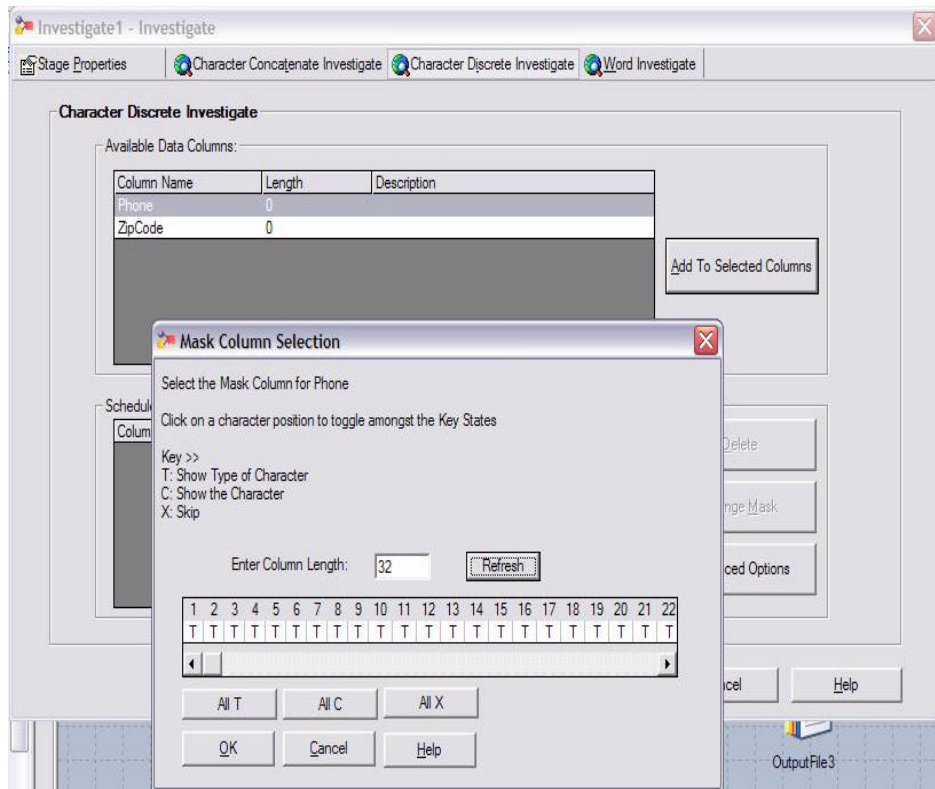


Figure 22-8 Investigate stage -> Character Discrete Investigate -> Add to Selected Columns button.

- c. As displayed in Figure22-8, enter a Column Length, and Click the Refresh button. (We entered a value of 32.)

You have the choice to display the T/Type of character, C/Character itself, or skip over (n) marked characters. Generally we always use T/Type.

Click OK when you are done.

- d. Repeat Steps.5-b,c above for the second/remaining column.

You are done with the step when both of our input columns have been added and configured.

- e. Move to the Stage Properties TAB -> Output TAB -> Mapping TAB.

Drag and drop all of the available output columns to the right side of this display. Example as shown in Figure22-9.

Click OK when your display equals that as shown in Figure22-9.

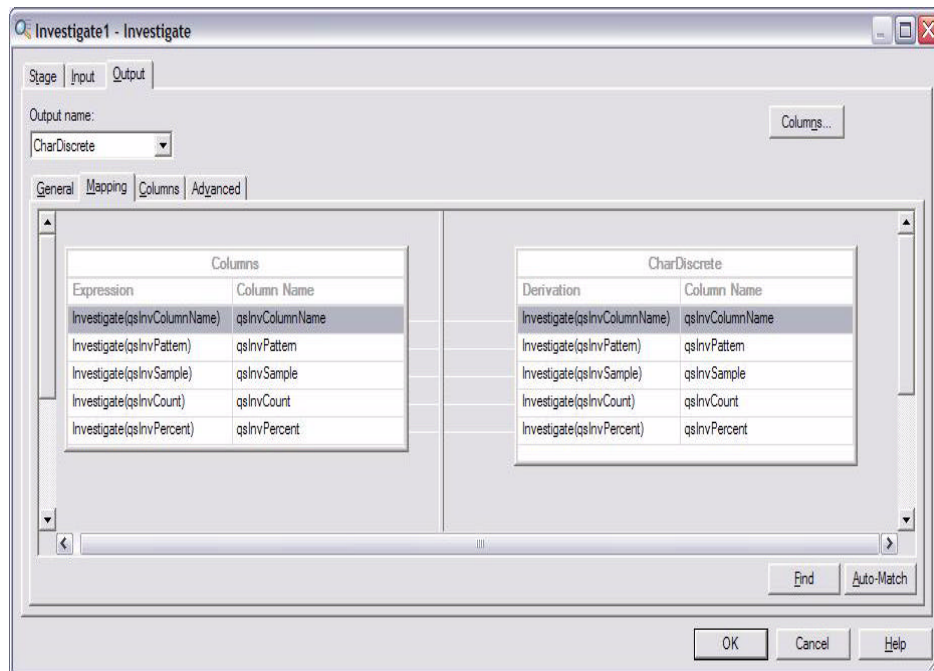


Figure 22-9 Investigate stage -> Stage Properties TAB -> Output TAB -> Mapping TAB.

- f. And click OK to exit and save the first Investigate stage properties dialog box.
6. Configure the second Investigate stage, the one for the Word Investigate reports (there are two reports available here)-
 - a. Double-click the Investigate stage to open its Properties dialog box.

Click the TAB entitled, Word Investigate.

If the preceding Copy stage is configured correctly, one column should be displayed as displayed in Figure22-10, (Company, or CompanyName, whatever you called your source input column).

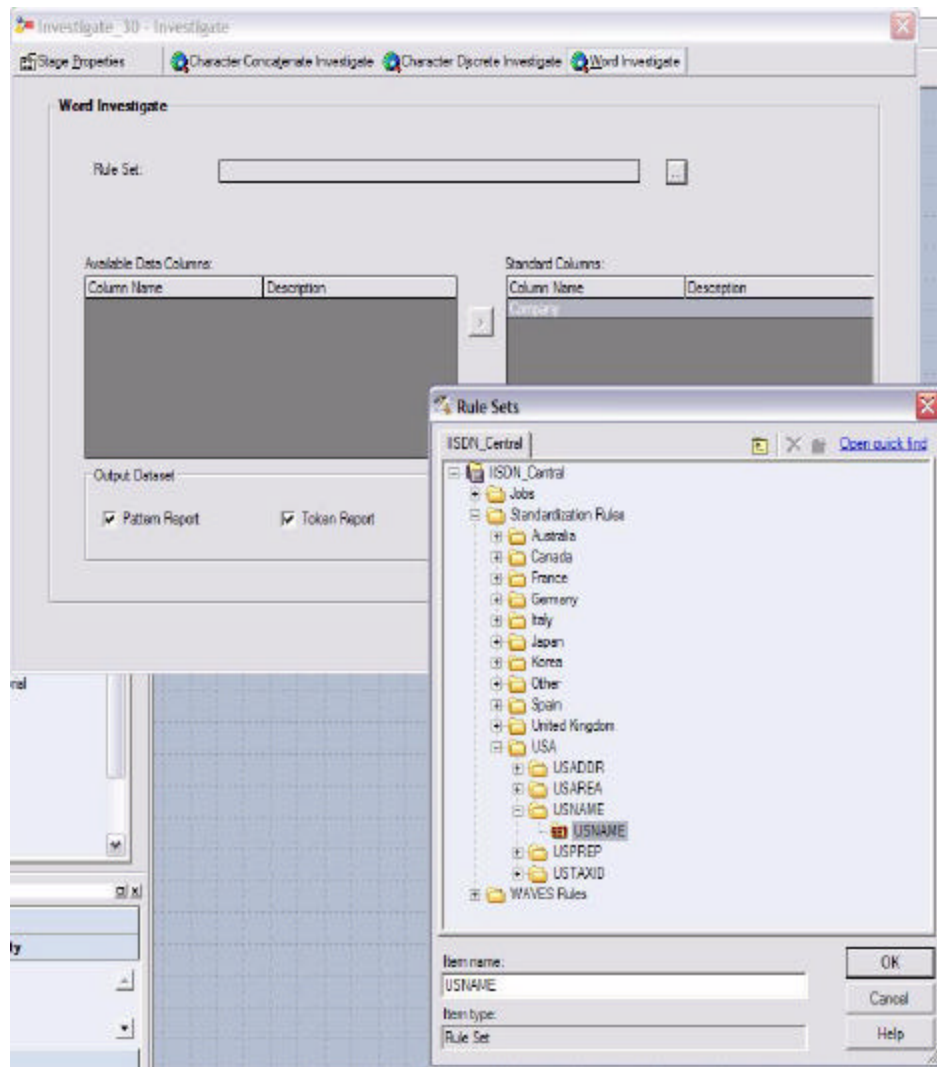


Figure 22-10 Investigate stage -> Word Investigate TAB.

- b. In Figure22-10, Click the Right-Arrow button to move the Company Name column from the left side of the display (Available Data Columns) to the right (Standard Columns).

- c. In the Output Dataset area of the display, Check both Pattern Report and the Token Report controls.
- d. Click the Ellipsis button “...” for Rule Set, and Browse to the USName rule under, Standardization Rules -> USA -> USNAME.
Click OK to Select USNAME. Example as shown in Figure 22-10.
- e. Move to the Stage Properties TAB -> Output TAB -> Mapping TAB, and call to output all available (output) columns. Example as shown in Figure 22-11.

You have to perform this step twice, because there are two output links to this (Investigate) stage. There is a drop down list box entitled, “Output name” that you must access.

The available output columns for each of these two links differ, because the output of these two distinct reports differ.

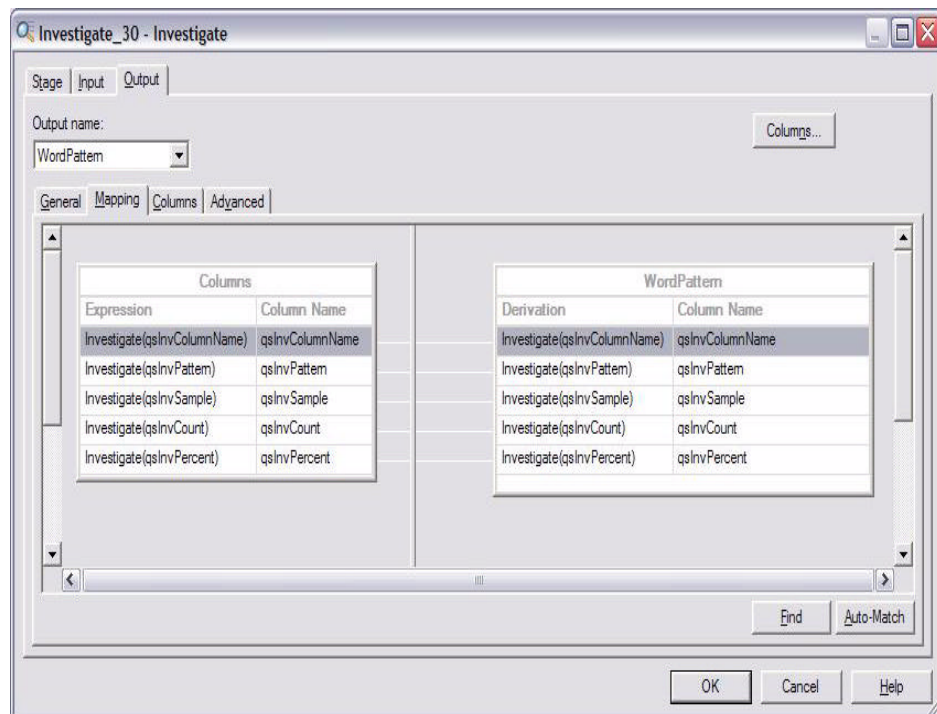


Figure 22-11 Investigate stage -> Stage Properties TAB -> Output TAB -> Mapping TAB.

7. Save, Compile, and Test.

That's it, you're done. Save, compile and test your Job.

A successful test appears equal to that as displayed in Figure22-3, and produces reports similar to those as displayed in Figure22-4, Figure22-5, and Figure22-6.

If you wish to experiment further, change the input text columns and applied rule set from Steps.6,b-e above, and test the USA -> USADDR rule set against address data. (Assuming you have address type data available.)

22.3 In this document, we reviewed or created:

We examined the (IBM) InfoSphere Information Server (IIS), QualityStage component, Investigate stage. Specifically, we used the Character Discrete report to view patterns in our input data sets, as a prelude to possibly creating any custom rule sets to standardize and cleanse this data. (This edition of IISDN gives focus to the Investigate stage. Next month we create custom rule sets which would benefit from this analysis.) Further, we used the Word Token and Word Pattern reports, for the same general purpose.

Persons who help this month.

Lady Dayo Joseph and future daughter Danielle, and Andy Wilson.

Additional resources:

The IBM InfoSphere Information Server, QualityStage component, product tutorial.

Legal statements:

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating trademarks that were owned by IBM at the time this information was published. A complete and current list of IBM trademarks is available on the Web at <http://www.ibm.com/legal/copytrade.shtml>

Other company, product or service names may be trademarks or service marks of others.

Special attributions:

The listed trademarks of the following companies require marking and attribution:

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.

Java and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Microsoft trademark guidelines

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Intel trademark information

UNIX is a registered trademark of The Open Group in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

ITIL is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office

IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency, which is now part of the Office of Government Commerce.

Other company, product, or service names may be trademarks or service marks of others.

