Smart Metadata in Cognos Insight 2012 Q3 release

Christine Fulford and Mohsen Rais-Ghasem
2012 July 19

Smart Metadata is the name of a set of new algorithms introduced in Cognos Insight in the Cognos 10.2 release.

When the user brings data into Cognos Insight, a model is created to enable analysis. Smart Metadata reduces the amount of modeling a user needs to do. Smart metadata looks at the column titles, a sample of the values, and the relationship of the columns to construct measures and dimensions. In many cases, Smart Metadata will create a sufficient model without any user modeling at all. Easier modeling leads to reduced complexity for the user and faster time to value.

Invocation

There are three ways to get data into Cognos Insight:

1. Drag data into the main window
2. Data | Quick Import
3. Data | Import Data

Smart Metadata is used in all three cases.

In some cases, the model is built without any user input at all. This happens if the user drags data in or does a “Quick Import” as long as the data is below a certain size (!!! what is this?).

Otherwise, a “guided import” is done. The guided import shows the import data. It allows the user to identify measures vs non-measures. And the Advanced dialog allows the user to build hierarchies.

Heuristic vs Deterministic

Smart Metadata and other Smart Coaches make extensive use of **heuristic** algorithms. Heuristic algorithms often refer to experience-based problem solving or learning techniques that are employed when deterministic approaches are expensive or not viable.

For example, Smart Metadata assumes that data items (e.g. columns in a spreadsheet) that are ‘related’ and should be grouped together (such Product Name, Product ID) tend to appear next to each other. This is not always true but to avoid a global search and mach algorithm that could be quite expensive, this 'rule of thumb' is acceptable.

Another example would be certain data clues such as presence of negative values or 1000 separator formatting to distinguish between category identifiers (such as 'product id') from measures (such as 'products sold'). Although not true deterministically but often category ids are non-negative and don't have such formatting.

Clues

Smart Metadata uses variety of clues to recognize and organizer the data items in a data source into a metadata model. The clues can be broadly grouped into two groups, lexical (i.e. column labels) and data (such as data types, formatting styles etc.).

The lexical clues often provide the best insight into what data items represent and hence are the primary source used by Smart Metadata. For example, the appearance of words such as ID, Num, Code, Key etc. in column headings such as 'ProductId', 'Product Code', or 'ENTRY_KEY' provides a strong evidence that the data contains keys or identifiers to some categories and should not be treated as a measure.

While very powerful, our lexical approach faces some challenges as well:

- It is limited to the extent of its vocabulary, which for the first release is very small (about 120 entries).
- Not very helpful in the absence of 'readable' column headings.
- Easily confused by lexical ambiguity, for example 'days' could mean day of the week (as a category) or a period of time (duration) as a measure.
- Our linguistic processing approach currently is keyword-based in a sense that words are examined individually not collectively, for example 'StudentNumber' is first broken into two words, 'student' and 'number' and examined individually; which is what will happen for 'Number of Students', as well.

Using lexical clues and other clues, Smart metadata helps with the detection of measures and various metadata hierarchies in Cognos Insighht.

Measure Detection

Smart Metadata looks for clues to distinguish measures (also known as metrics) from non-measures. Only numeric values are identified as measures. However, it can be tricky to distinguish numeric non-measures (like Purchase Number) from measures (like Number of Purchases). The kinds of clues Smart Metadata uses to distinguish numeric non-measures from measures are

- The presence of non-numeric characters. For example "(202) 123-4567" is readily identified as a non-measure.
- The presence of a decimal character (usually a period but this depends on the locale). Values like 23.87 usually indicate a measure. (But not always - for example IP addresses have periods in them).
- Column titles containing the word ID or Code suggest a non-measure
- A uniform number of digits often indicate a code or identifier

If your data includes a measure that Cognos Insight isn't recognizing, here are some ideas to consider

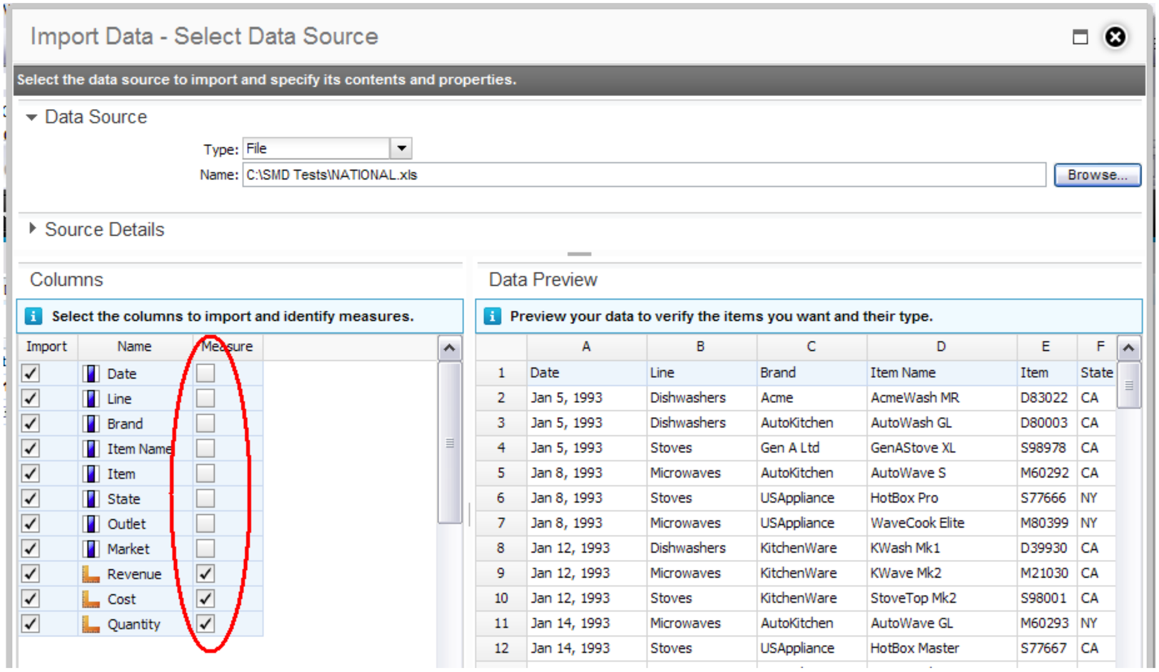
- Check the column title to see if it might be confused with some other concept. For example, if the title has the word Day or Month or Country or State or Code or ID in it, Cognos Insight is less likely to consider it a measure. Rename the column if that is reasonable.
- Check for non-numeric values and non-numeric characters in the data

If your data includes a non-measure that Cognos Insight recognizes as a measure, here are some ideas to consider.

- Add the word Code or ID to the end of the column title.
- If many rows of data contain a period, see if the period can be removed without changing the meaning of the data. For example, if there is a .0 at the end of every value, remove it.

For example, imagine a column titled "Release" with values like 8.4, 8.5, 9.1, 9.2. Changing the column title to "Release ID" might be enough to get Cognos Insight to recognize it as a non-measure. If not, another approach might be to change the values to 8-4, 8-5, 9-1, 9-2.

If none of these ideas are helpful or practical and Cognos Insight continues to incorrectly interpret a column as a measure or non-measure, you can change this in Cognos Insight. Use the Import Data… option in the Data menu to bring in your data, rather than Quick Import or simply dragging the data in. The first dialog has checkboxes for each column in your data source that indicate whether it's a measure or not. Simply toggle the setting.



Import Data - Select Data Source

Select the data source to import and specify its contents and properties.

Data Source

Type: File

Name: C:\SMD Tests\NATIONAL.xls

Source Details

Columns

Select the columns to import and identify measures.

Import	Name	Measure
☑	Date	☐
☑	Line	☐
☑	Brand	☐
☑	Item Name	☐
☑	Item	☐
☑	State	☐
☑	Outlet	☐
☑	Market	☐
☑	Revenue	☑
☑	Cost	☑
☑	Quantity	☑

Data Preview

Preview your data to verify the items you want and their type.

	A	B	C	D	E	F
1	Date	Line	Brand	Item Name	Item	State
2	Jan 5, 1993	Dishwashers	Acme	AcmeWash MR	D83022	CA
3	Jan 5, 1993	Dishwashers	AutoKitchen	AutoWash GL	D80003	CA
4	Jan 5, 1993	Stoves	Gen A Ltd	GenAStove XL	S98978	CA
5	Jan 8, 1993	Microwaves	AutoKitchen	AutoWave S	M60292	CA
6	Jan 8, 1993	Stoves	USAppliance	HotBox Pro	S77666	NY
7	Jan 8, 1993	Microwaves	USAppliance	WaveCook Elite	M80399	NY
8	Jan 12, 1993	Dishwashers	KitchenWare	KWash Mk1	D39930	CA
9	Jan 12, 1993	Microwaves	KitchenWare	KWave Mk2	M21030	CA
10	Jan 12, 1993	Stoves	KitchenWare	StoveTop Mk2	S98001	CA
11	Jan 14, 1993	Microwaves	AutoKitchen	AutoWave GL	M60293	NY
12	Jan 14, 1993	Stoves	USAppliance	HotBox Master	S77667	CA

Time Hierarchy Detection

Smart Metadata looks for clues in the data to identify a time dimension. These clues include

- Column titles containing time-related words like Date, Year, Month, Week, Day.
- Column data that is clearly and consistently dates.
- Appropriate data values. For example Month values from 1 to 12

If your data includes a time dimension that Cognos Insight doesn't recognize, here are some ideas to consider

- Use a meaningful time-related column title. While it's likely to recognize a column with title "Year" as being time-related, it's less likely to recognize one with title "Y"
- Check the data format. If it's a date, it's most recognizable if it uses a date format like those used in Excel. If it's a year, Smart Metadata will be more readily identify values like "2001, 2002, 2003" as being time-related and will be less certain about values like "1, 2, 3" or year values with comma separators like "2,012".
- Scan the column of data for non-date values. For example, a last name accidentally placed in the date column will reduce the likelihood of the column being correctly identified as time-related.

- Scan the column of data for out-of-range values. For example, a value greater than 12 in a “Month” column reduces the likelihood of the column being correctly identified as time-related.

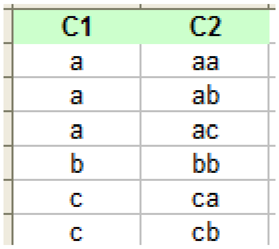
Geographical Hierarchy Detection

Smart Metadata recognizes some common geographical entities such as city, country, state/province and organize them in proper order, such as Country → State → City. It also has some knowledge of generic geo. entities such as region, county, and so on that are organized, along with the known entities, in a single hierarchy based on their cardinalities (i.e. number of unique values). The items with smaller cardinality are to appear before (‘higher’ level) than the items with higher cardinality.

Detection of well-known Business Hierarchies

Smart Metadata also recognizers certain well-known business hierarchies, namely Product hierarchy (e.g. Line → Brand → Item), customer hierarchy (e.g. customer group → customer), industry (sector type → sector) and Sales organization. This determination is driven largely by lexical clues, and the cardinality of the data items (for the level order). However the hierarchy members are also expected to demonstrate a one-to-many association in the data.

This determination is done by sampling the data (up to first 2000 rows) and ensuring that each value on from the first column is associated with a unique set of values on the second column. For example, given the sample data below one can assume C1 values have a one-to-many association with V2 values



C1	C2
a	aa
a	ab
a	ac
b	bb
c	ca
c	cb

Multi-Lingual Support

Smart Metadata works in both English and other languages. They keywords that Smart Metadata recognizes have been translated.

Also for a selected languages such as English, German, and French, some provisions have been made to recognize word variants. For example, Smart Metadata recognizes both Day and its plural form Days as relating to the same concept. This will not work as well for all languages that support plurals. A general rule that works for many European languages has been integrated, but we don't have language-specific rules nor do we manage the thousands of exceptions.

Many data input files in other countries are in English or use English terms like "ID". When Cognos Insight is working in other languages, it first tries to match against the translated terms and then it tries the English terms.

Conclusion

Smart Metadata is a set of heuristic algorithms in Cognos Insight that contribute to forming a model. Smart Metadata provides a faster start to meaningful analysis. Since Smart Metadata uses heuristic algorithms, it won't always get it right. Knowledge of the factors Smart Metadata takes into account can be used in creating input files that work well with Smart Metadata. Also, in many cases, guided data import can be used to correct any wrong choices Smart Metadata made and augment the model with the user's additional knowledge.