

Data Profiling & Exploration with Pentaho Data Integration

Bridging the gap between data and insight by leveraging analytics in the data integration process

Will Munji

Solution Architect, Enterprise Architecture Group

April-2018

Overview & Use Cases

Data Explorer Views

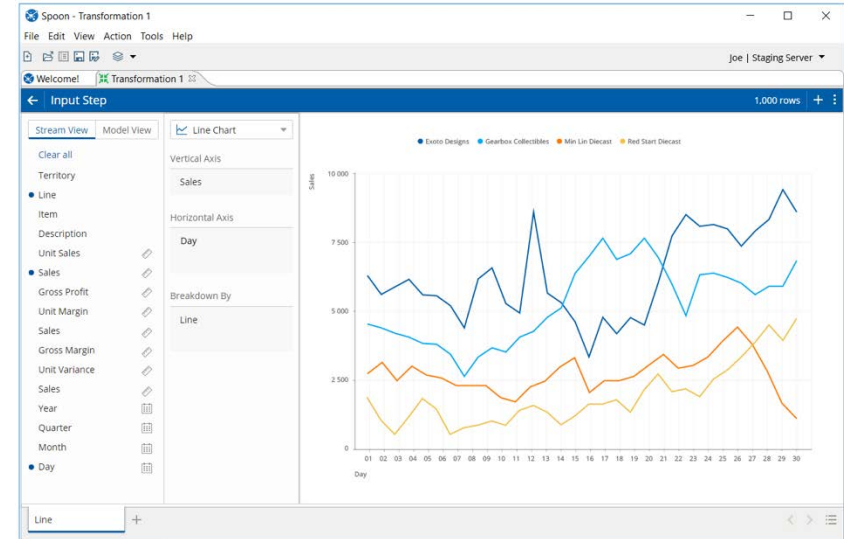
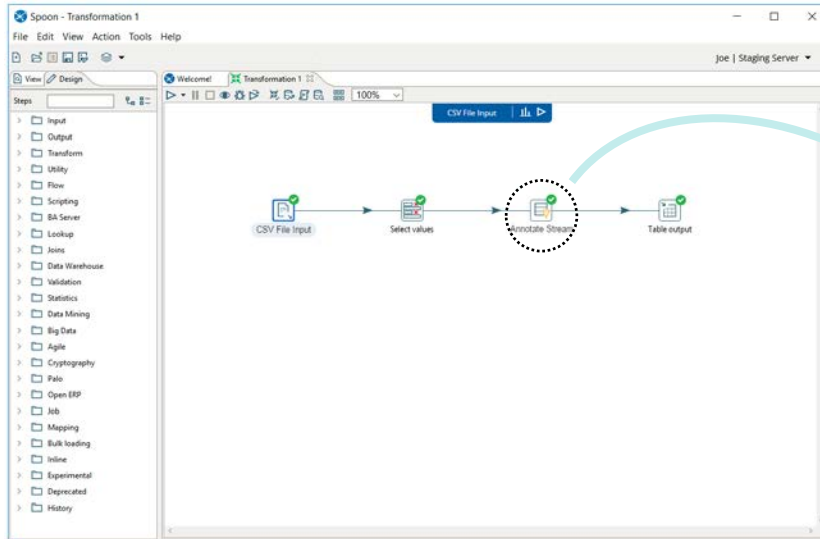
Filtering Data in Data Explorer

Some Usage Considerations

Demos

Overview & Use Cases

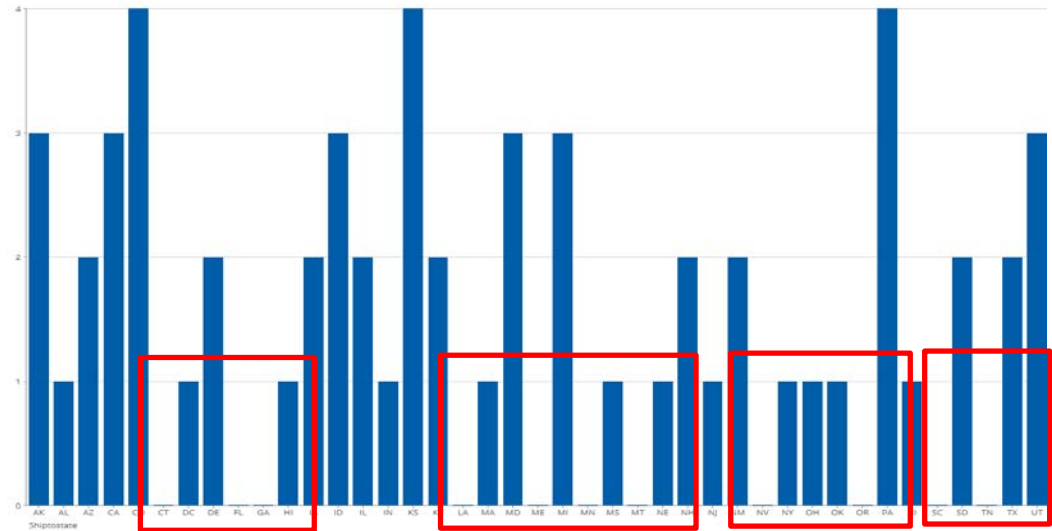
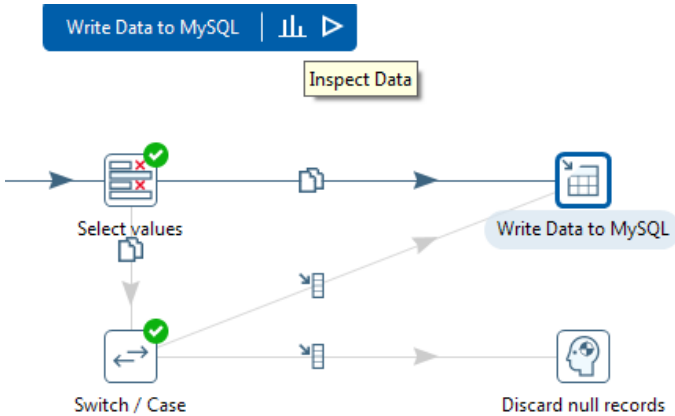
Data Explorer in PDI



**Access visualizations during data prep for inspection or prototyping –
and accelerate time to insight**

Use Case – Data Inspection

Identify missing or incorrect data during the data prep process.

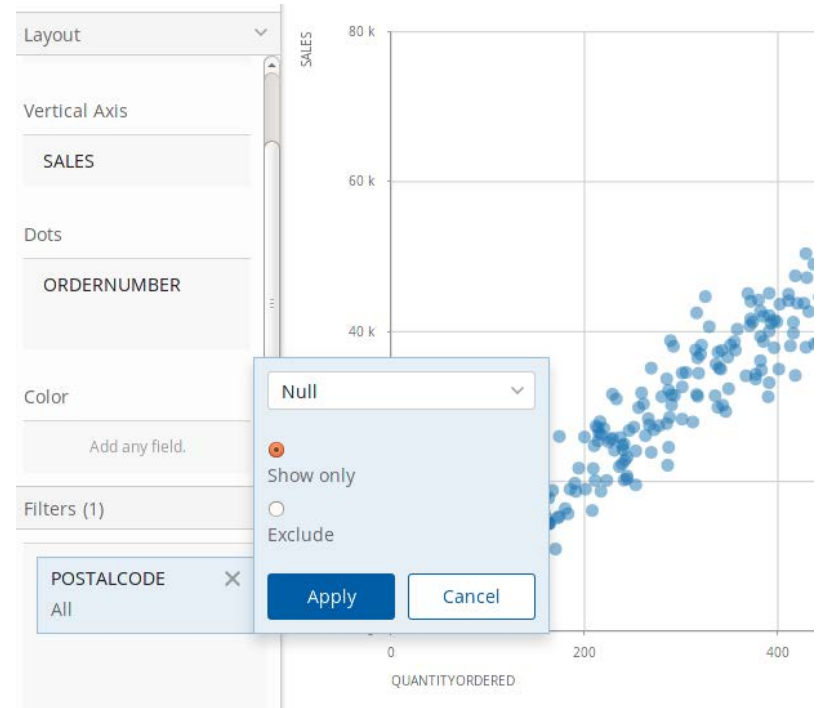




Use Case – Data Inspection Cont'd

Filter data on-the-fly

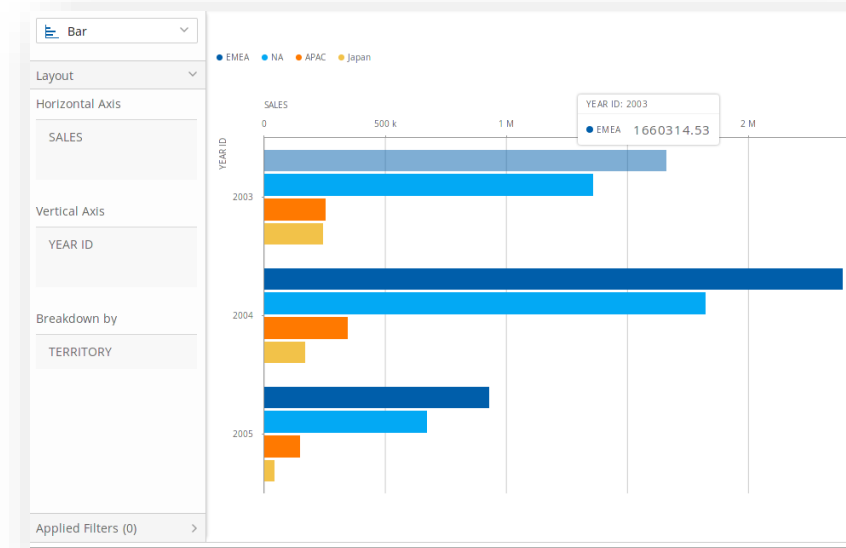
- Apply restrictions to include/exclude certain data when using charts in Data Explorer
- Filters can be applied to numeric and non-numeric fields
- Examples: State contains 'California', Sales > 1000, Address is NOT Null, Exclude England



Use Case – BI Prototyping

Model and visualize

- Model data on-the-fly
- Detect or annotate **hierarchies**
- Quickly apply visualizations to data
- **Drill-down** to lower levels on charts and pivot tables



Use Case – BI Prototyping Cont'd

Spoon - Transformation 1

File Edit View Action Tools Help

Joe | Staging Server

Input Step

Stream View Model View

Table

1,000 rows +

Re-run the transformation
Exit without stopping
Publish Data Source

T...	Line	Item	Description	Un...	Sales	Gros...	Unit...	Sales	Sales	Sales	Sales
APAC	Feed	Alfalfa	50lb Alfalfa square bale	25,458	154,154	54,124	25,458	154,154	54,124	25,458	154,154
APAC	Feed	Bahia	50lb Bahia square bale	23,145	187,654	98,542	23,145	187,654	98,542	23,145	187,654
APAC	Feed	Bermuda	50lb Bermuda square bale	22,447	154,963	65,487	22,447	154,963	65,487	22,447	154,963
APAC	Feed	Coastal	50lb Coastal square bale	26,654	154,154	95,324	26,654	154,154	95,324	26,654	154,154
APAC	Feed	Orchard	50lb Orchard square bale	23,145	187,654	65,347	23,145	187,654	65,347	23,145	187,654
APAC	Feed	Timothy	50lb Timothy square bale	22,447	154,963	54,124	22,447	154,963	54,124	22,447	154,963
APAC	Feed	Wheat	40lb bag whole wheat	12,985	154,154	98,542	12,985	154,154	98,542	12,985	154,154
APAC	Feed	Oats	40lb bag whole oats	13,457	187,654	65,487	13,457	187,654	65,487	13,457	187,654
APAC	Feed	Mixed Grains	40lb bag corn, oats, sunfl...	11,674	154,963	95,324	11,674	154,963	95,324	11,674	154,963
APAC	Feed										
APAC	Feed										
APAC	Feed										
APAC	Hardware										
APAC	Hardware										
APAC	Hardware										
APAC	Hardware	4" Hog Panel	4 foot hog panel	31,458	154,963	65,487	12,124	154,963	65,487	31,458	154,963
APAC	Hardware	1/4" Hardware cloth	6ft x 100ft welded wire	451	154,154	65,487	451	154,154	65,487	451	154,154
APAC	Hardware	4" Hog Panel	4 foot hog panel	31,458	154,963	65,487	12,124	154,963	65,487	31,458	154,963

Publish data sources from PDI directly to business analytics tools.

Spoon - DET_main_demo

File Edit View Action Tools Help

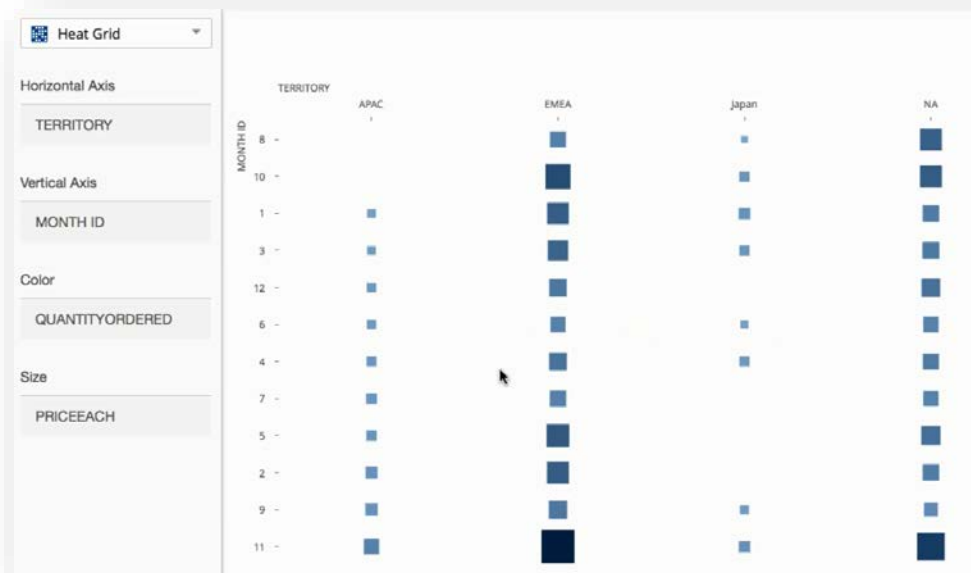
Connect

Publish Data Source

Create a data source for users on the Pentaho Server.

Help Get Started Close

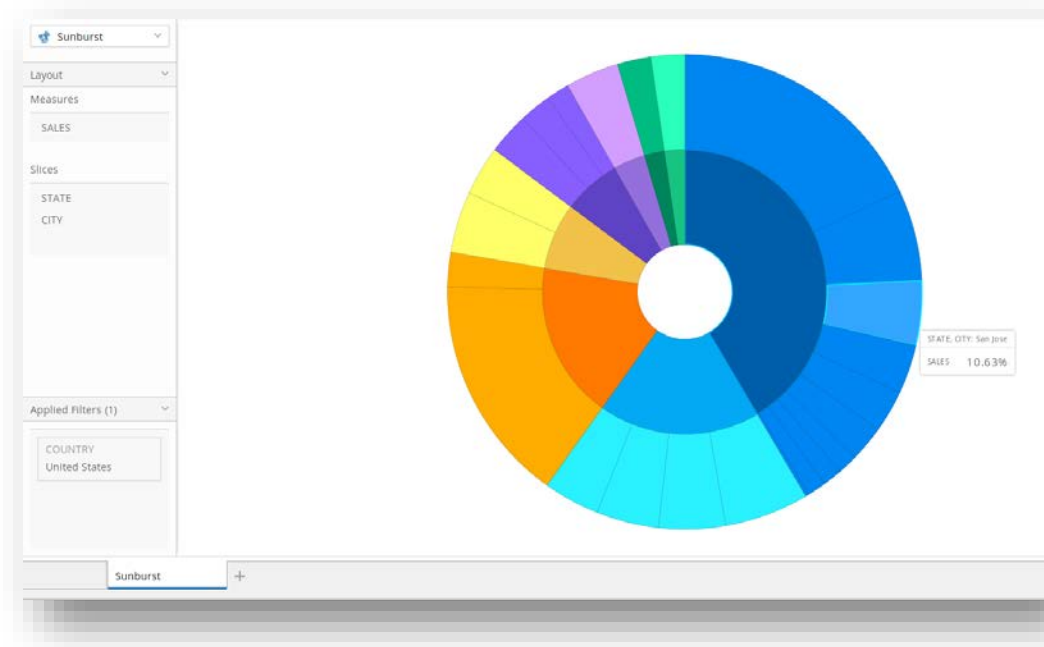
Heat Grid Visualization



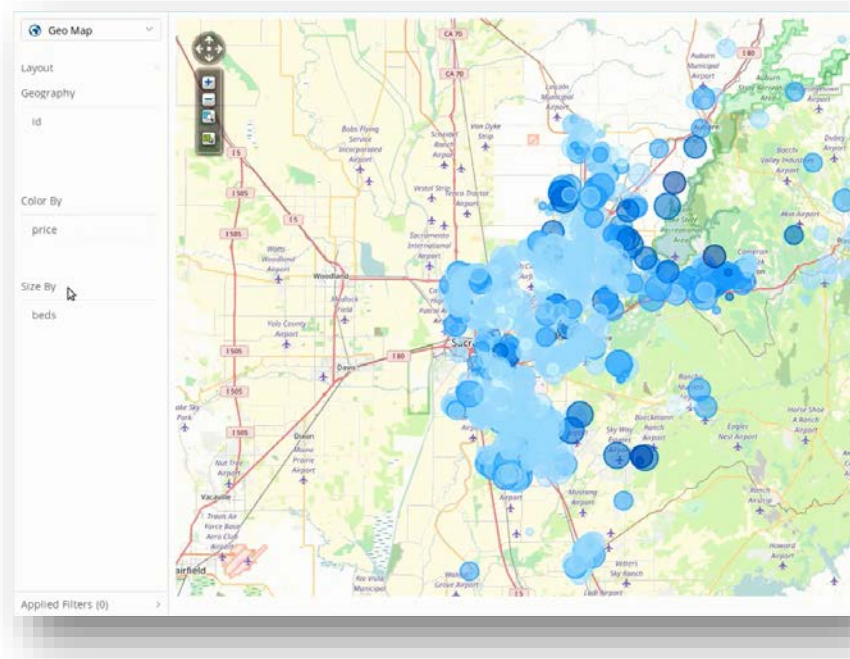
- Similar to Analyzer chart, shows 2 dimensions and 2 measures at once
- Dimensions are on axes, color and size of points vary by measure value
- Most useful for relative comparisons at the 'intersection' of 2 dimensions
 - Ex: See sales metrics by each combination of month and region (as shown)

Sunburst Visualization

- Similar to Analyzer, useful for showing how a measure is distributed across several categories / attributes
 - Esp. useful for showing multiple levels in hierarchy at once
- Ex: breakdown of sales by state (inner slice), and city (outer slice)



Geo Map Visualization



- Similar to Analyzer, measures represented by dot size/color
 - Pan, zoom actions
- Same Auto-geocoding as Analyzer
 - Auto plot for: lat/lng, certain countries, their subdivisions, their cities, US county/zip

Data Explorer Views

Stream and Model Views

Stream View

Model View

Spoon - Transformation 1
File Edit View Action Tools Help
Joe | Staging Server
Welcome! Transformation 1
Input Step 1,000 rows +

Stream View Model View

Table

T...	Line	Item	Description	Un...	Sales	Gros...	Unit...	Sales	Re-run the transformation Exit without stopping Publish Data Source		
APAC	Feed	Alfalfa	50lb Alfalfa square bale	25,458	154,154	54,124	25,458	154,154	54,124	25,458	154,154
APAC	Feed	Bahia	50lb Bahia square bale	23,145	187,654	98,542	23,145	187,654	98,542	23,145	187,654
APAC	Feed	Bermuda	50lb Bermuda square bale	22,447	154,963	65,487	22,447	154,963	65,487	22,447	154,963
APAC	Feed	Coastal	50lb Coastal square bale	26,654	154,154	95,324	26,654	154,154	95,324	26,654	154,154
APAC	Feed	Orchard	50lb Orchard square bale	23,145	187,654	65,347	23,145	187,654	65,347	23,145	187,654
APAC	Feed	Timothy	50lb Timothy square bale	22,447	154,963	54,124	22,447	154,963	54,124	22,447	154,963
APAC	Feed	Wheat	40lb bag whole wheat	12,985	154,154	98,542	12,985	154,154	98,542	12,985	154,154
APAC	Feed	Oats	40lb bag whole oats	13,457	187,654	65,487	13,457	187,654	65,487	13,457	187,654
APAC	Feed	Mixed Grains	40lb bag corn, oats, sunfl...	11,674	154,963	95,324	11,674	154,963	95,324	11,674	154,963
APAC	Feed	Layer Pellets	20lb Purina Layena	31,145	154,154	65,347	31,145	154,154	65,347	31,145	154,154
APAC	Feed	Grower Pellets	20lb Purina Flock Raiser	21,145	187,654	54,124	21,145	187,654	54,124	21,145	187,654
APAC	Feed	Starter Pellets	20lb Purina Fock Starter	22,124	154,963	98,542	22,124	154,963	98,542	22,124	154,963
APAC	Hardware	1/4" Hardware cloth	6ft x 100ft welded wire	451	154,154	65,487	451	154,154	65,487	451	154,154
APAC	Hardware	4" Hog Panel	4 foot hog panel	31,458	154,963	65,487	12,124	154,963	65,487	954	154,963
APAC	Hardware	1/4" Hardware cloth	6ft x 100ft welded wire	451	154,154	65,487	451	154,154	65,487	451	154,154
APAC	Hardware	4" Hog Panel	4 foot hog panel	31,458	154,963	65,487	12,124	154,963	65,487	954	154,963
APAC	Hardware	1/4" Hardware cloth	6ft x 100ft welded wire	451	154,154	65,487	451	154,154	65,487	451	154,154
APAC	Hardware	4" Hog Panel	4 foot hog panel	31,458	154,963	65,487	12,124	154,963	65,487	954	154,963

Table +

Stream View

- No modeling layer used, just SQL
- Uses PDI data types and masks
- Required for flat table



Field Name	Icon
Zipcode	123
State	
City	
Year	123
Month	123
Day	123
Hour	123
Totalonlinesales	123
Totalsessioncount	123
Retailstoresalesa...	123
Totalscannedeven...	123
Country	

Model View

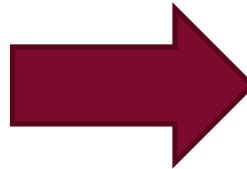
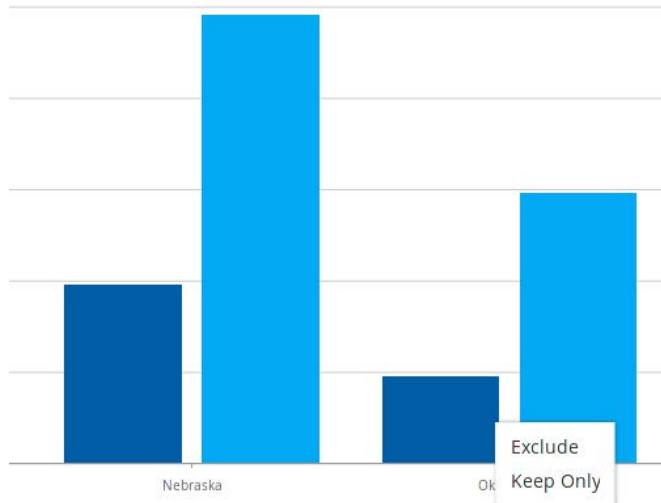
- Uses Measures and Attributes specified in BA model layer
- Required for pivot table, geo map, and sunburst charts



Section	Item Name	Icon
Measures	retailstoresalesam...	123
	totalonlinesales	123
	totalscannedevent...	123
	totalsessioncount	123
Time	year	123
	month	123
	day	123
	hour	123

Stream View – “Drill Through” Scenario

- Visualize the shape of the data (select any chart)
- apply non-numeric filters to narrow focus and
- switch back to table view to see the underlying records for granular inspection.



Customer	Country	State C...	State	Claim A...	Response
YH23384	US	NE	Nebraska	2412.8	Yes
KO26461	US	NE	Nebraska	3227	No
OU78470	US	NE	Nebraska	2932.8	No
UI64281	US	OK	Oklahoma	2285.6	No
QL59704	US	NE	Nebraska	2344.5	No
HB67642	US	NE	Nebraska	3461.1	Yes
NV61299	US	NE	Nebraska	2779	Yes
MY37953	US	NE	Nebraska	2583.1	No
TN36521	US	NE	Nebraska	2498	No
FA46418	US	OK	Oklahoma	2471	No

Filtering Data in Data Explorer

Filters Pane

The screenshot displays a data application interface. On the left is a 'Filters Pane' with a search bar and a list of filterable fields: Customer, Country, State Code, State, Claim Amount, Response, Coverage, Education, Effective To Date, EmploymentStatus, Gender, Income, Location Code, Marital Status, Monthly Premium, Months Since Last..., Months Since Poll..., and Number of... The 'Filters (1)' section is highlighted with a red box. On the right is a data table with columns: Customer, Country, State C..., State, and Claim A... A modal dialog is open over the table, showing a filter configuration for 'Income' with the operator 'Does not equal' and the value '0'. The dialog has 'Apply' and 'Cancel' buttons.

Customer	Country	State C...	State	Claim A...
BU79786	US	KS	Kansas	276.4
QZ44356	US	NE	Nebraska	698
AI49188	US	OK	Oklahoma	1288.7
WW63253	US	MO	Missouri	764.6
HB64268	US	KS	Kansas	281.4
OC83172	US	IA	Iowa	825.6
XZ87318	US	IA	Iowa	538.1
CF85061	US	NE	Nebraska	721.6
DY87989	US	IA	Iowa	2412.8
BQ94931	US	IA	Iowa	738.8
SX51350	US	MO	Missouri	473.9
VQ65197	US	MO	Missouri	819.7
DP39365	US	MO	Missouri	879.9
SJ95423	US	NE	Nebraska	881.9
WJ66600	US	MO	Missouri	538.4
			Iowa	746.3
			Oklahoma	256.7
			Missouri	394.5
			Iowa	571
			Missouri	816.3
			Iowa	287.2
			Kansas	304.2
YH23384	US	NE	Nebraska	2412.8
TZ98966	US	OK	Oklahoma	245

- Drag and drop onto the filters pane
- Filters can be edited from filters pane

Options from Filters Pane – Numeric Fields

Greater Than / Less Than

Effective To Date
EmploymentStatus
Gender
Income
Filters (3)
Claim Amount X
> 3000

Greater than

or equal to

Value
3000

Apply Cancel

Equals / Does Not Equal

Policy
Contains "Personal"

Income X
≠ 0

Does not equal

Value
0

Apply Cancel

Null

Income
≠ 0

Total Claim A... X
Exclude Nulls

Null

Show only
 Exclude

Apply Cancel

***NOTE** – in Model View, there is no Null filter for Measures

Options from Filters Pane – Non-Numeric Fields

Equals / Does Not Equal

A screenshot of a filter configuration dialog. The dialog has a title bar with a close button (X). The main area contains a dropdown menu set to 'Equals' and a text input field containing 'NJ'. Below the input field are 'Apply' and 'Cancel' buttons. The background shows a list of filters with 'State = NJ' selected.

Null

A screenshot of a filter configuration dialog. The dialog has a title bar with a close button (X). The main area contains a dropdown menu set to 'Null' and two radio buttons: 'Show only' (unselected) and 'Exclude' (selected). Below the radio buttons are 'Apply' and 'Cancel' buttons. The background shows a list of filters with 'City Exclude Nulls' selected.

***Note** – These filters match on a certain string; there is no 'pick from list' filter as in Analyzer

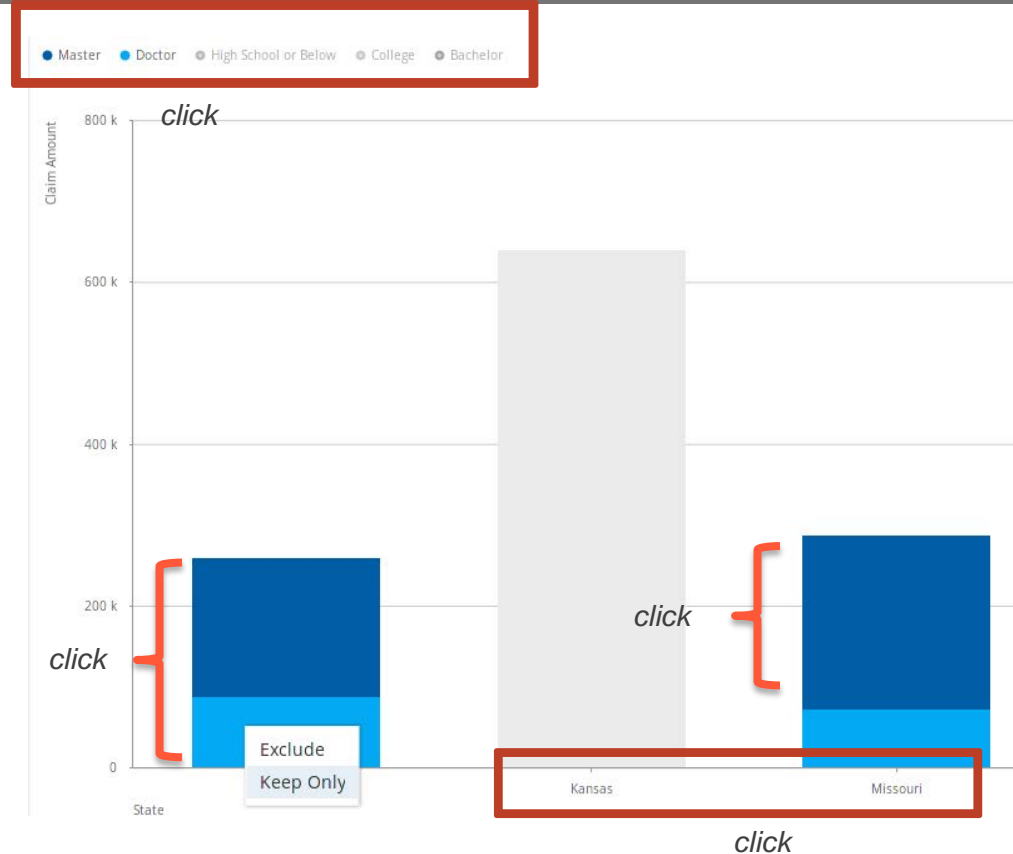
Contains / Does Not Contain

A screenshot of a filter configuration dialog. The dialog has a title bar with a close button (X). The main area contains a dropdown menu set to 'Contains' and a text input field containing 'Personal'. Below the input field are 'Apply' and 'Cancel' buttons. The background shows a list of filters with 'Policy Contains "Personal"' selected.



Chart Actions

- Create filters in charts and perform drill-down actions
- Chart segments, legends, and labels are all clickable for filtering
- Multi-select requires holding ctrl key + click
- Cannot edit these filters once created (only remove)



Demos

Insurance Claim Data Explorer

- Explore insurance claim data
- Publicly-available data on prediction website, Kaggle
- Simulate how a data scientist could use PDI to quickly analyze and visualize data



kaggle

NYPD Motor Vehicle Collisions

- Explore motor vehicle collisions
- Publicly-available data on NYC Open Data
- Simulate how a data scientist could use PDI to quickly visualize data and project on a map



NYC OpenData

- What we covered today:
 - Background on Data Explorer (DE) and its main use cases – inspection / data prep and BI Prototyping
 - Deeper dive on specific DE features and how to use them – visualizing, modeling, filtering, publishing, and more
 - Demonstration of DE in action

- Want to learn more?
 - For documentation on DE, search “*Inspect Data*” on **help.pentaho.com**
 - This webinar, slides and other videos will be available online

Questions?

