HITACHI
Inspire the Next

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

Agenda

HITACHI
Inspire the Next

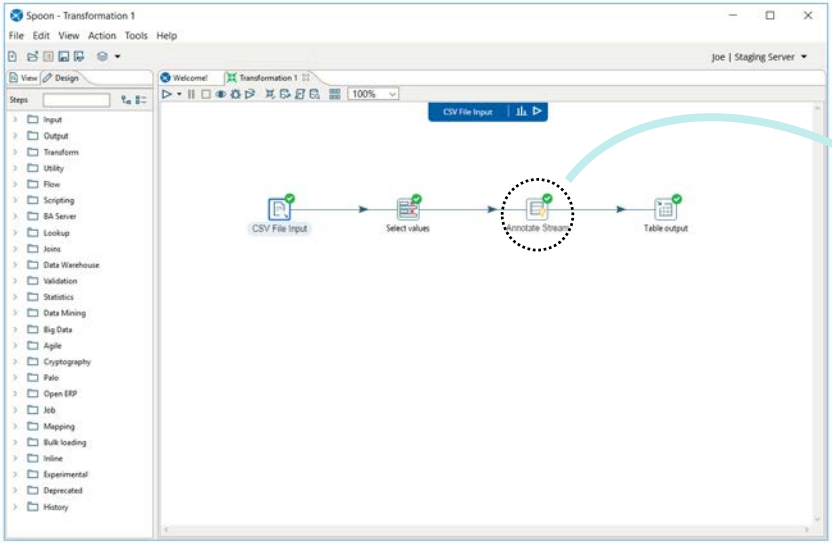
- Overview & Use Cases
- Data Explorer Views
- Filtering Data in Data Explorer
- Some Usage Considerations
- Demos

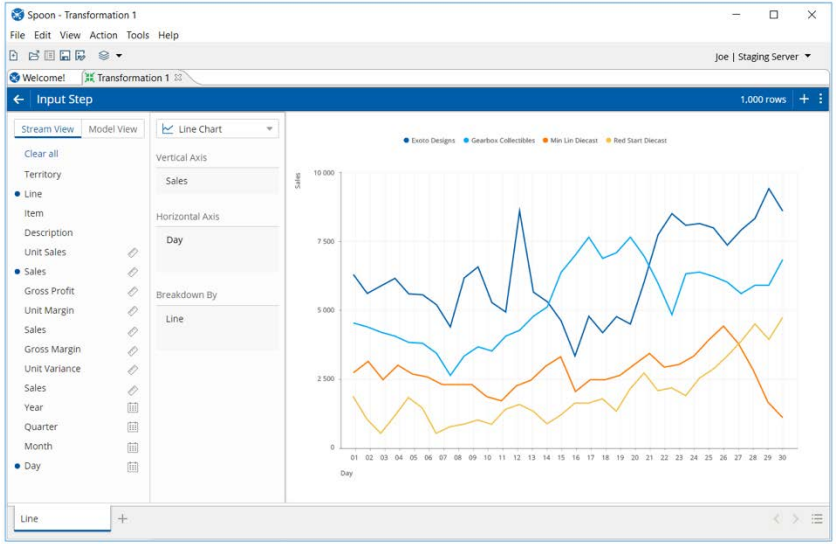
HITACHI
Inspire the Next

Overview & Use Cases

Data Explorer in PDI

HITACHI
Inspire the Next



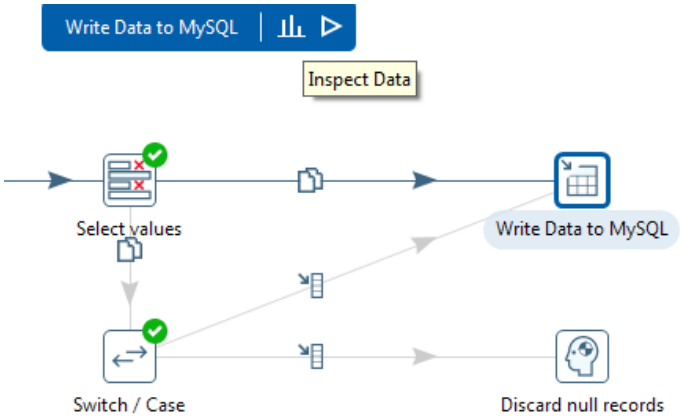


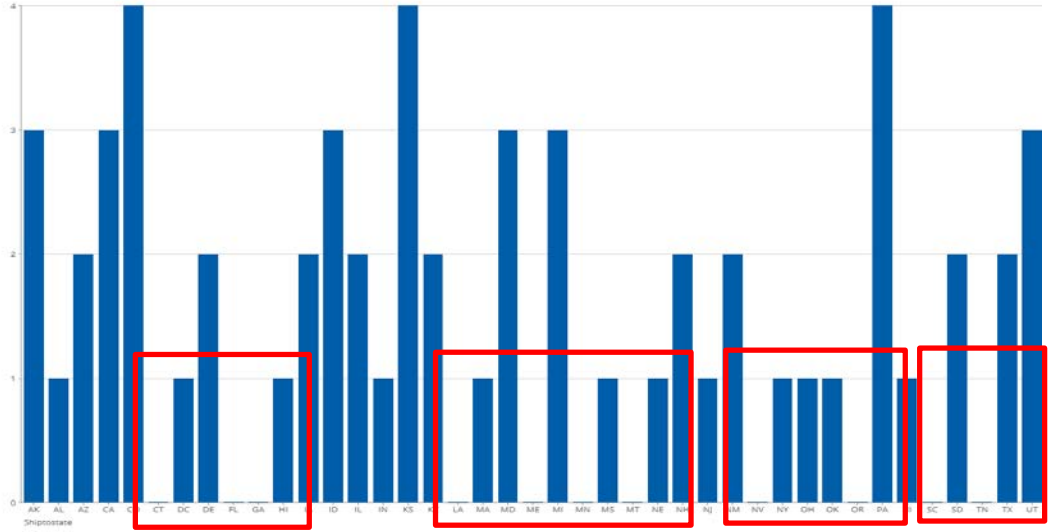
Access visualizations during data prep for inspection or prototyping – and accelerate time to insight

Use Case – Data Inspection

HITACHI
Inspire the Next

Identify missing or incorrect data during the data prep process.





Use Case – Data Inspection Cont'd

HITACHI
Inspire the Next

Filter data on-the-fly

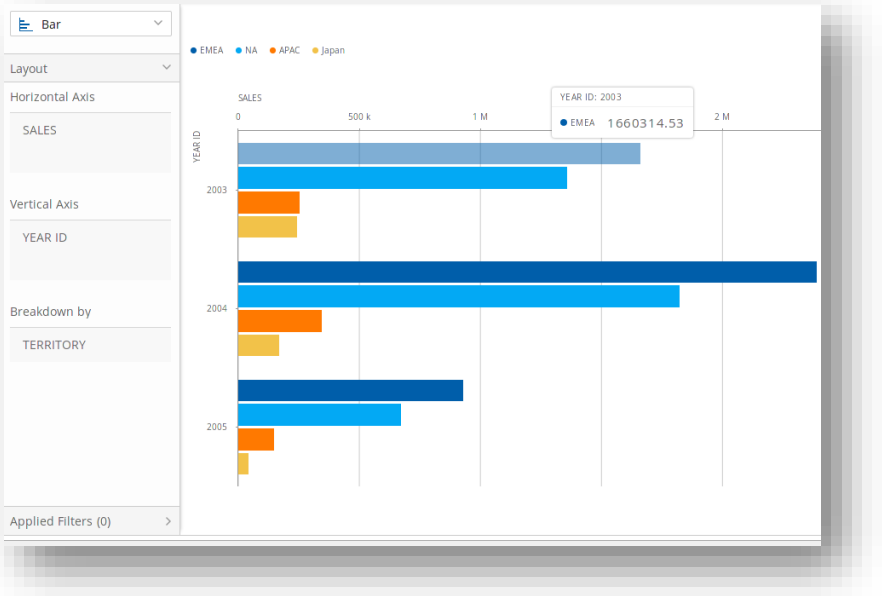
- Apply restrictions to include/exclude certain data when using charts in Data Explorer
- Filters can be applied to numeric and non-numeric fields
- Examples: State contains 'California', Sales > 1000, Address is NOT Null, Exclude England

Use Case – BI Prototyping

HITACHI
Inspire the Next

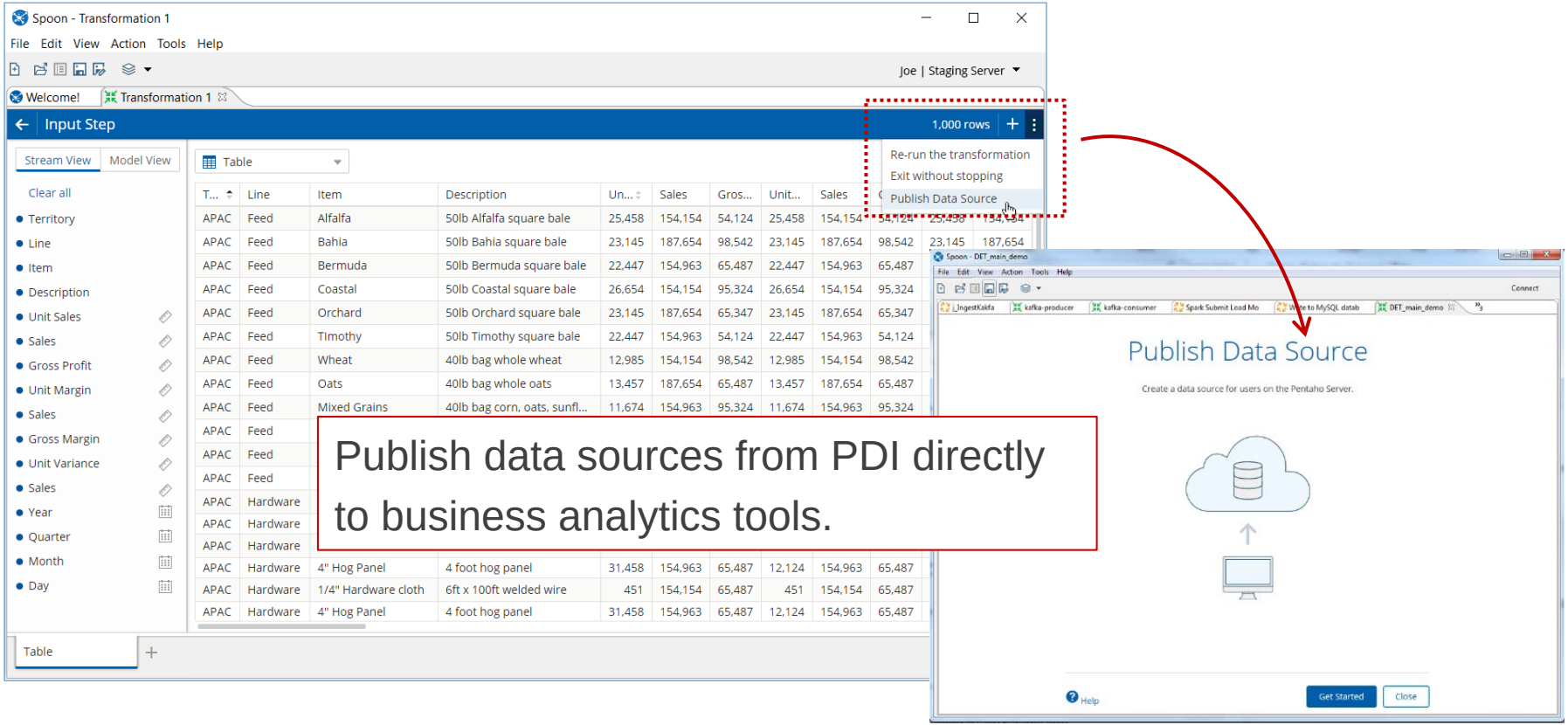
Model and visualize

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



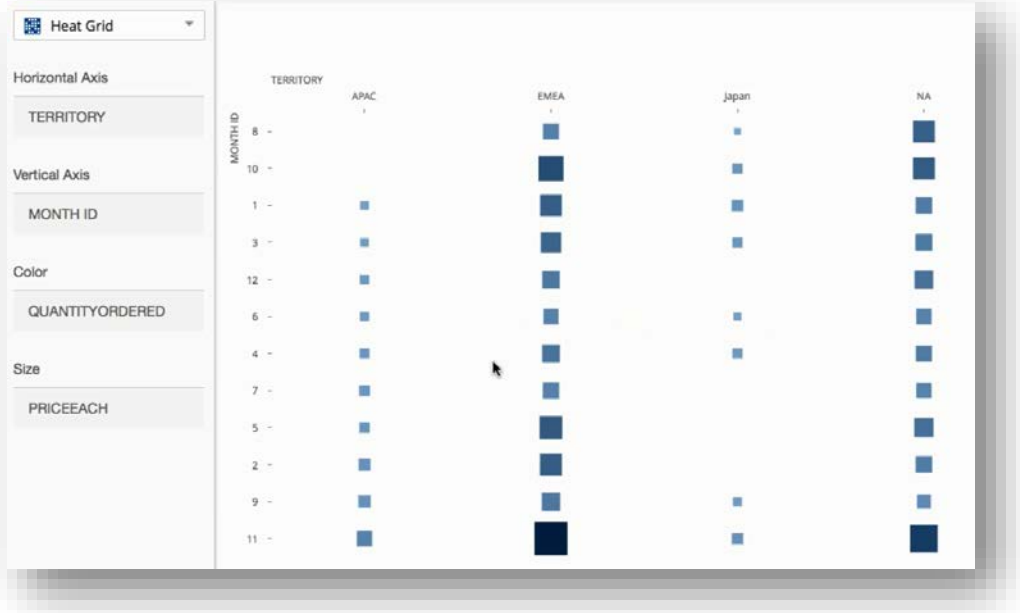
Use Case – BI Prototyping Cont'd

HITACHI
Inspire the Next



Publish data sources from PDI directly to business analytics tools.

Heat Grid Visualization

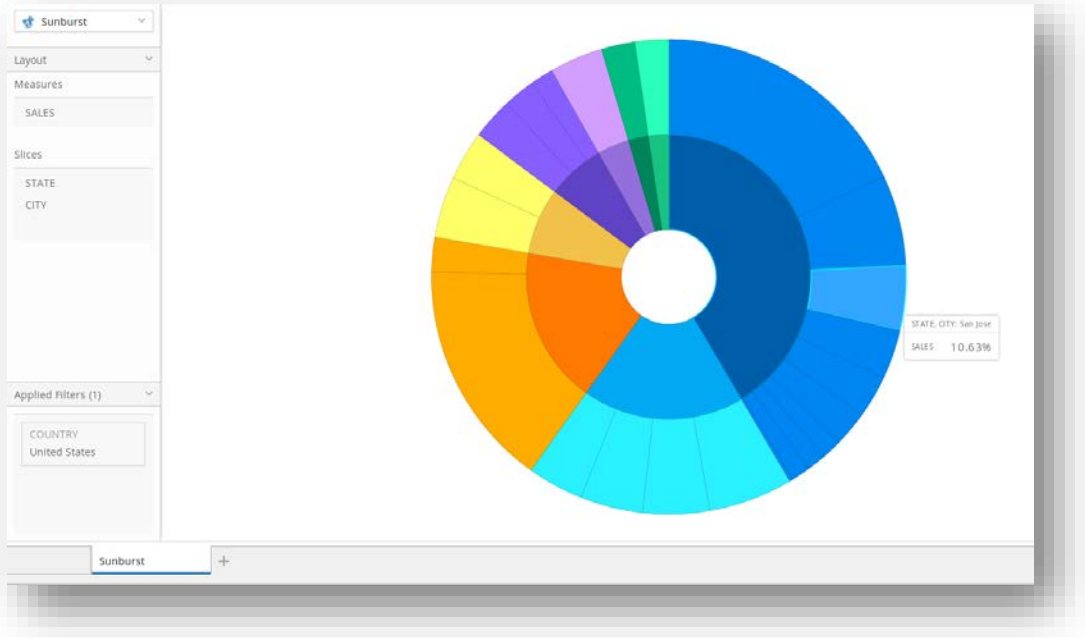


- Similar to Analyzer chart, shows 2 dimensions and 2 measures at once
- Dimensions are on axes, color and size of points vary by measure value
- Most useful for relative comparisons at the ‘intersection’ of 2 dimensions
  - Ex: See sales metrics by each combination of month and region (as shown)

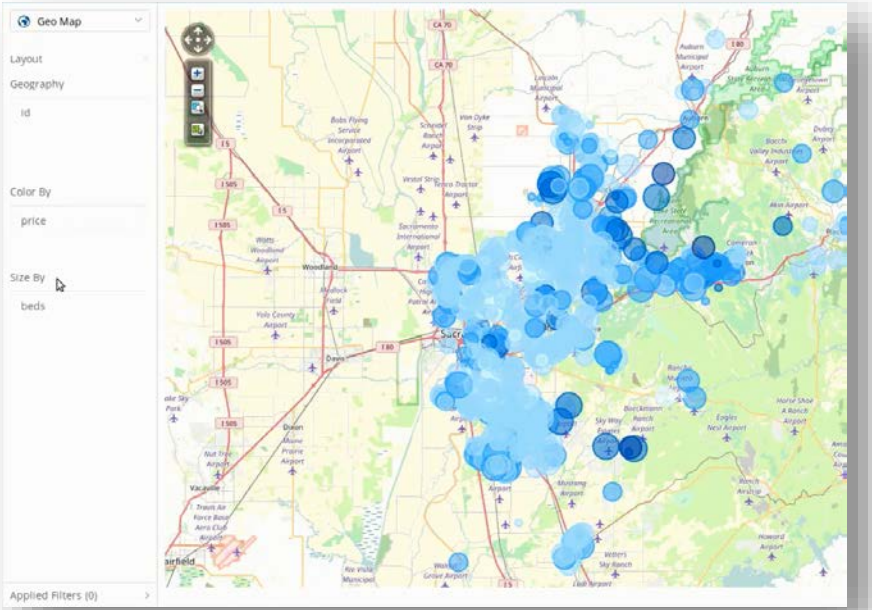
Sunburst Visualization

HITACHI
Inspire the Next

- 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



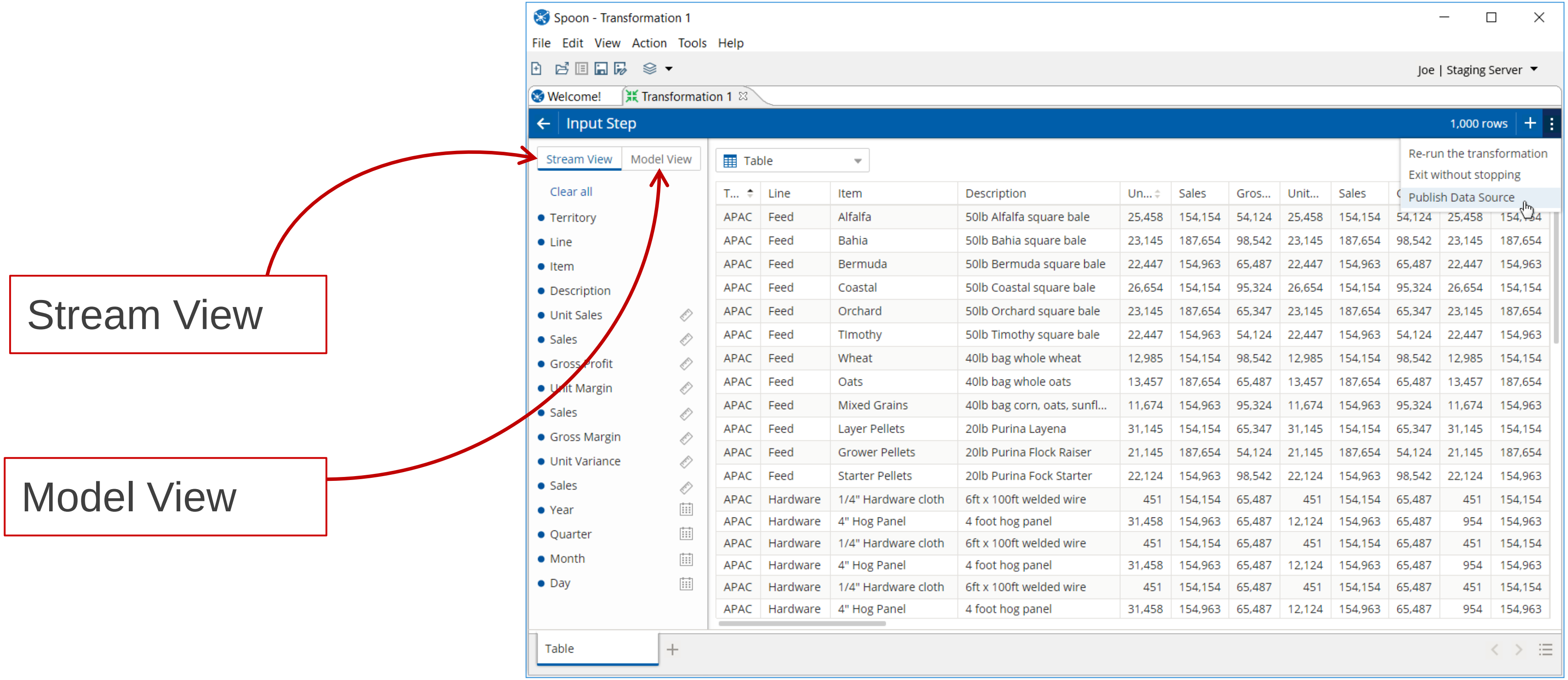
- Similar to Analyzer, measures represented by dot size/color
  - Pan, zoom actions
- Same Auto-geocoding as Analyzer
  - Auto plot for: lat/lng, certain countries, their subdivisions, their cities, US county/zip

HITACHI
Inspire the Next

Data Explorer Views

Stream and Model Views

HITACHI
Inspire the Next

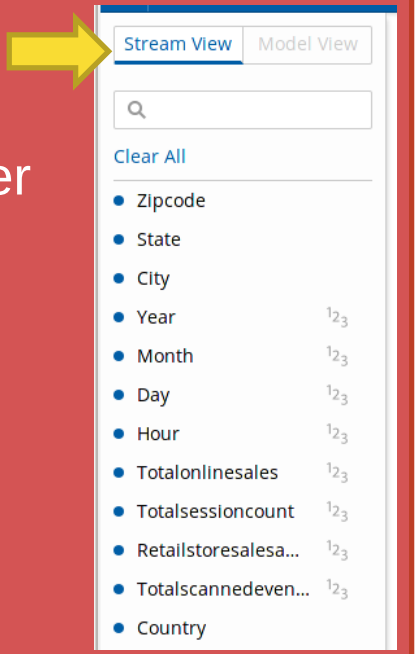


Stream and Model Views Cont'd

HITACHI
Inspire the Next

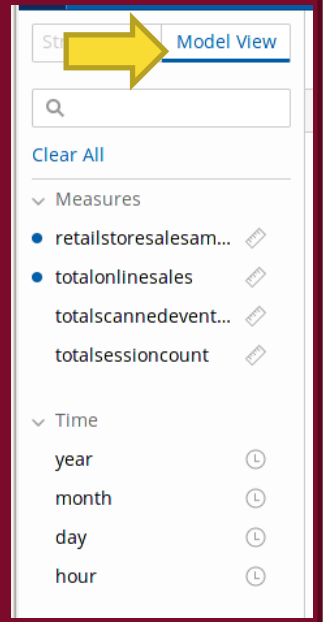
Stream View

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



Model View

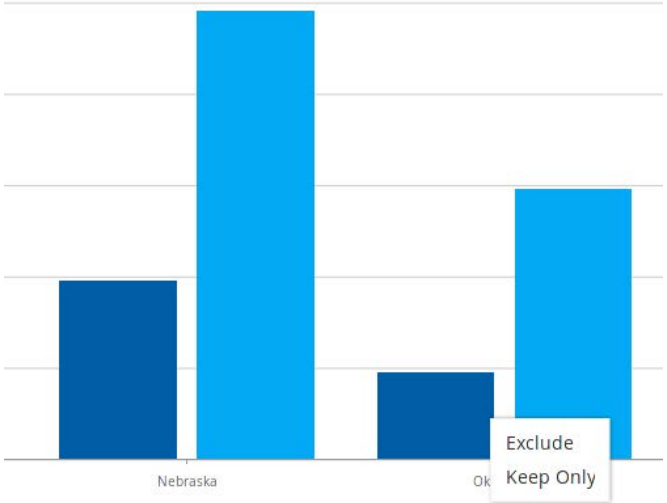
- Uses Measures and Attributes specified in BA model layer
- Required for pivot table, geo map, and sunburst charts



Stream View – *“Drill Through”* Scenario

HITACHI
Inspire the Next

- 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.

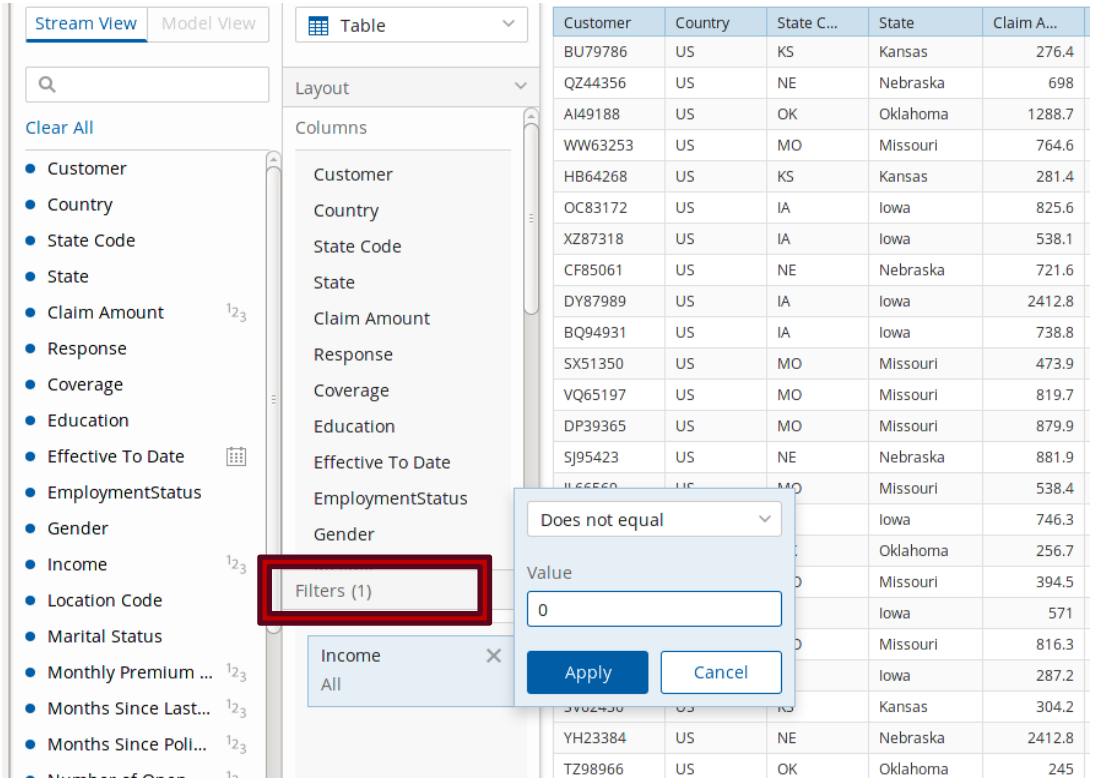


Customer	Country	State C...	State	Claim A...	Response
YH23384	US	NE	Nebraska	2412.8	Yes
KO26461	US	NE	Nebraska	3227	No
OU78470	US	NE	Nebraska	2932.8	No
UI64281	US	OK	Oklahoma	2285.6	No
QL59704	US	NE	Nebraska	2344.5	No
HB67642	US	NE	Nebraska	3461.1	Yes
NV61299	US	NE	Nebraska	2779	Yes
MY37953	US	NE	Nebraska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

HITACHI
Inspire the Next

Filtering Data in Data Explorer

Filters Pane

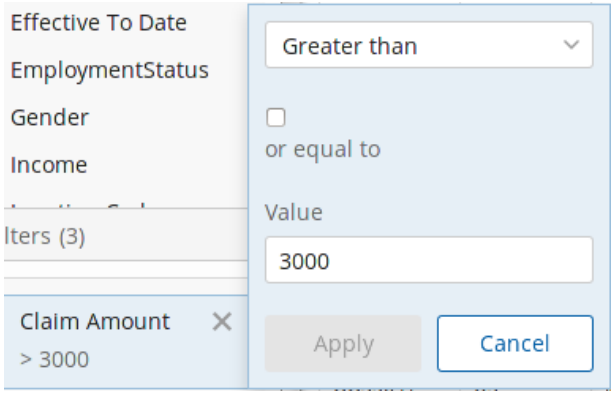


- Drag and drop onto the filters pane
- Filters can be edited from filters pane

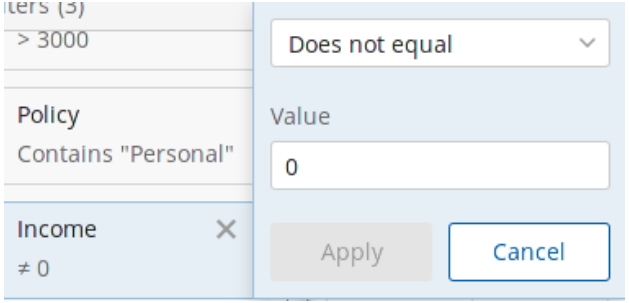
Options from Filters Pane – Numeric Fields

HITACHI
Inspire the Next

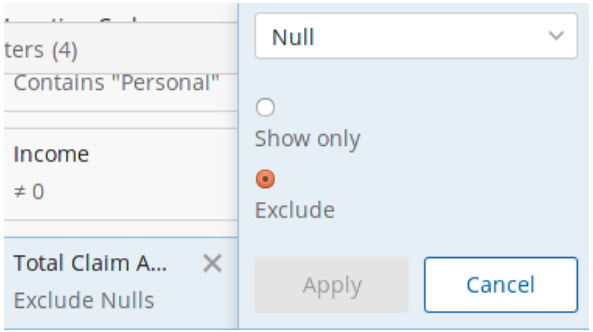
Greater Than / Less Than



Equals / Does Not Equal



Null

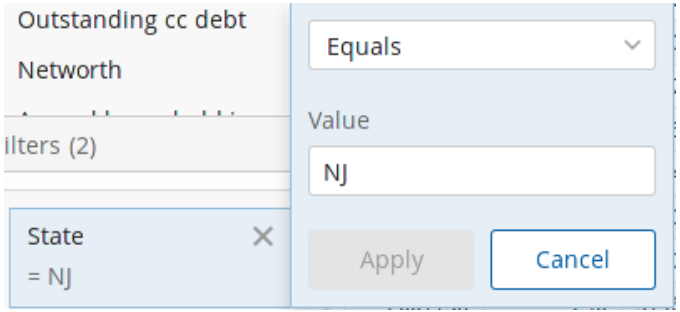


*__NOTE__ – in Model View, there is no Null filter for Measures

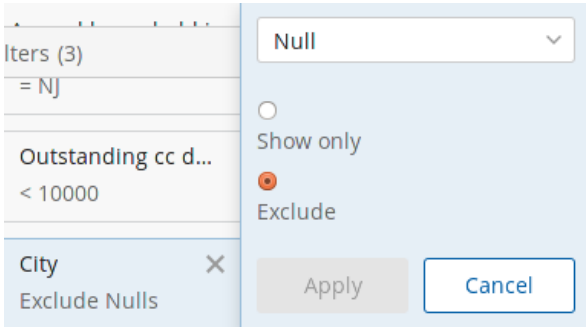
Options from Filters Pane – Non-Numeric Fields

HITACHI
Inspire the Next

Equals / Does Not Equal



Null



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

Contains / Does Not Contain

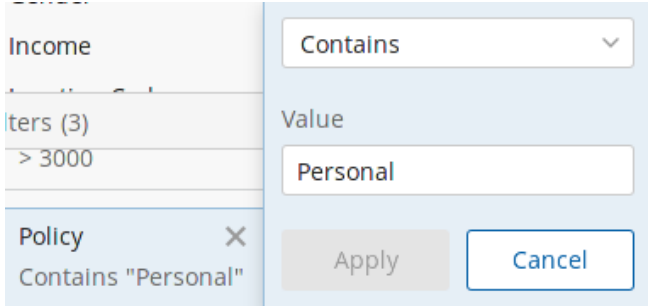


Chart Actions

- 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)

HITACHI
Inspire the Next

Demos

Insurance Claim Data Explorer

HITACHI
Inspire the Next

- 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

HITACHI
Inspire the Next

- 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

Summary

HITACHI
Inspire the Next

- 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

Next Steps

HITACHI
Inspire the Next

- 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?