HITACHI
Inspire the Next

Data Explorer in Pentaho Data Integration (PDI)

HITACHI

Inspire the Next

Change log (if you want to use it):

Date	Version	Author	Changes

Contents

This page intentionally left blank.

Overview

This document covers some best practices on using Data Explorer in Pentaho Data Integration (PDI) to quickly visualize and analyze data. Visual data exploration provides access to analytics during data preparation, letting you easily spot-check data issues without switching in and out of tools or waiting until the very end to find data quality problems. In addition, different departments can collaborate and iterate faster, shortening the cycle from raw data to meaningful analytics.

Our intended audience is Pentaho administrators, data analysts, or anyone with a background in PDI who is interested in configuring Data Explorer and using it to create analytics within the data preparation and integration phase of development.

The intention of this document is to speak about topics generally; however, these are the specific versions covered here:

Software	Version(s)
Pentaho	7.x, 8.0+

The Components Reference in Pentaho Documentation has a complete list of supported software and hardware.

Before You Begin

Before beginning, use the following information to prepare for the procedures described in the main section of the document.

Terms You Should Know

Here are some terms you should be familiar with:

- **Pentaho Data Integration (PDI):** provides access to an Extraction, Transformation, and Loading (ETL) engine that captures the right data, cleanses the data, and stores data using a uniform format that is accessible and relevant to end users and IoT technologies.
- **Pentaho Analyzer:** easy-to-use, web-based, drag-and-drop report design tool that can be used to create analytics against multi-dimensional data using online analytical processing (OLAP).
- **Pentaho User Console (PUC):** web-based console for managing users and content on the Pentaho platform.

Other Prerequisites

This document assumes that you are familiar with creating jobs and transformations in PDI and (optionally) Pentaho Server and have installed it in your environment.

© Hitachi Vantara Corporation 2018. All Rights Reserved

Example Use Cases

Here are a couple of use cases to keep in mind while working through this document:

Use Case 1: Business Intelligence Prototyping

Janice is a Pentaho administrator. Her organization uses PDI for developing data pipelines that ingest data into corporate data stores.

Janice wants to speed up development and shorten business review cycles by creating inflight visualizations within transformations, and prototyping models and reports. She has decided to try Data Explorer to achieve these goals.

Use Case 2: Quick Data Inspections

Fabiola is a member of the data science team within a technology company. She and her team usually spend hours and occasionally days ingesting, formatting and otherwise preparing data for predictive model construction.

Since Fabiola has learned that Data Explorer can be used to quickly visually inspect the data to determine outliers and overall patterns, she has suggested using this tool for her team's modeling objectives.

© Hitachi Vantara Corporation 2018. All Rights Reserved

Data Explorer

Data Explorer is a feature in PDI that enables data engineers, data scientists and business users to easily access visualizations during data preparation, for the purposes of inspection and prototyping, and to accelerate time to insights.

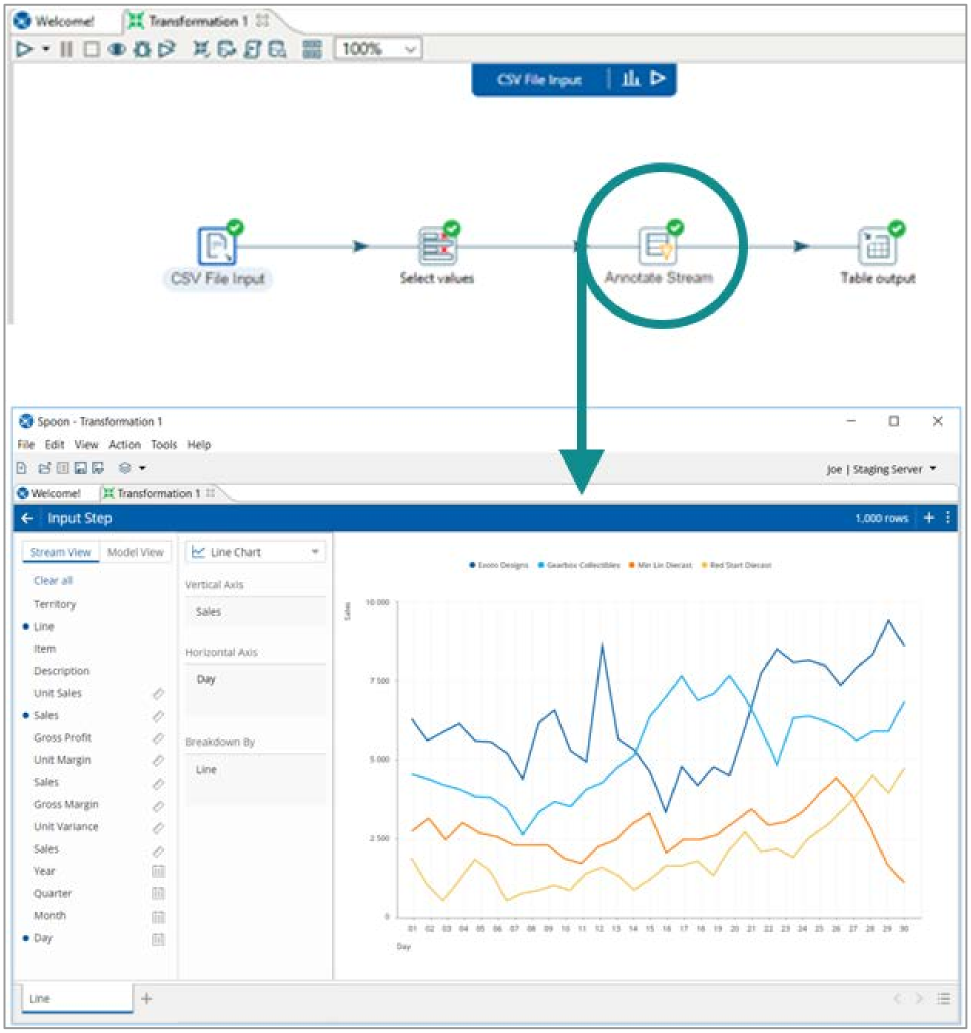


Figure 1: Data Explorer

You can find details on these topics in the following sections:

- Accessing Data Explorer
- Changing Maximum Number of Analyzed Records
- Opening Data Explorer on New Transformations
- Using Annotate Stream to Improve Model Results
- Drill-Down on Annotate Stream Data
- Filtering Null Measures (Numeric Fields)
- Filtering Aggregations in Tables or Charts
- Filtering Limitations

© Hitachi Vantara Corporation 2018. All Rights Reserved

Accessing Data Explorer

You can access Data Explorer from virtually every step in a PDI transformation by right-clicking and selecting **Inspect Data** from the context menu:

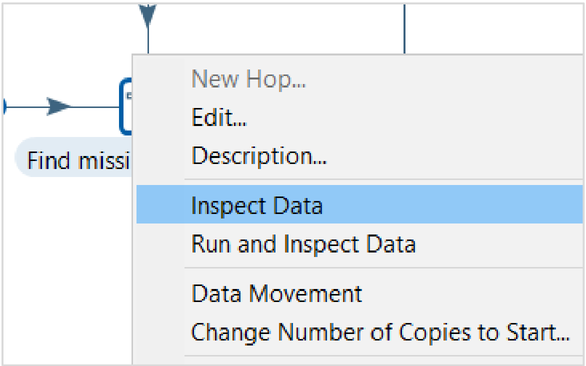


Figure 2: Inspect Data

Alternatively, access it by clicking on the **Inspect Data** button that appears on the top of the PDI screen when a step is selected. More information is available at Inspect Your Data.

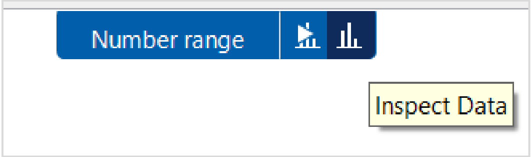


Figure 3: Inspect Data Popup

Changing Maximum Number of Analyzed Records

The default maximum number of records that Data Explorer reads for presentation and analysis is 50,000. Although you can increase this number, note that doing so may lead to performance issues.

Follow these steps to change the limit:

1. Open PDI.
2. Click **Edit** and then **Edit `kettle.properties`**.
3. Scroll down and create a new row.
4. For the **variable name**, use `det.dataservice.dynamic.limit`.
5. For the **value** field, enter any number other than `50000`.
6. The **description** field is optional.
7. Restart PDI.

© Hitachi Vantara Corporation 2018. All Rights Reserved

Opening Data Explorer on New Transformations

When you open Data Explorer, it loads rows incrementally until all records are loaded. While records are loading, a **processing** message will appear on the bottom of the screen.

Once all the records are completely loaded, subsequent opens take less time because the Data Explorer browser remains open in the background.

Using Annotate Stream to Improve Model Results

Using the **Annotate Stream** PDI step can improve Data Explorer's auto-model results, required to do business analytics (BA) data source prototyping. **Annotate Stream** allows for specification of the following types of objects:

- Time hierarchy
- Product hierarchy
- Specified measure formats/aggregations

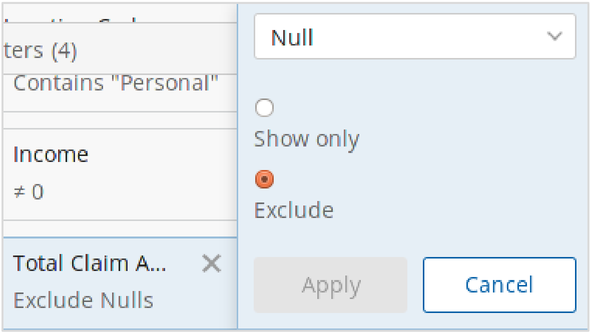
Drill-Down on Annotate Stream Data

Drill-down behavior (double-click action) on data processed through the **Annotate Stream** step is different from the drill action behavior on data processed through all other steps in PDI (such as **Select Values**, **Text File Input**, and others). The order of fields drilled when using **Annotate Stream** is as follows:

- **Bar/Column chart**: Drills first on breakdown field, then axis.
- **Scatter**: Drills first on color field, then dots.
- **Heat grid**: Drills first on horizontal field, then vertical.
- **Sunburst**: Drills based on the slice you click.

Filtering Null Measures (Numeric Fields)

In the model view, there is no `null` filter for measures. Use **Exclude**:



ters (4)

Contains "Personal"

Income

≠ 0

Total Claim A...

Exclude Nulls

Null

○ Show only

● Exclude

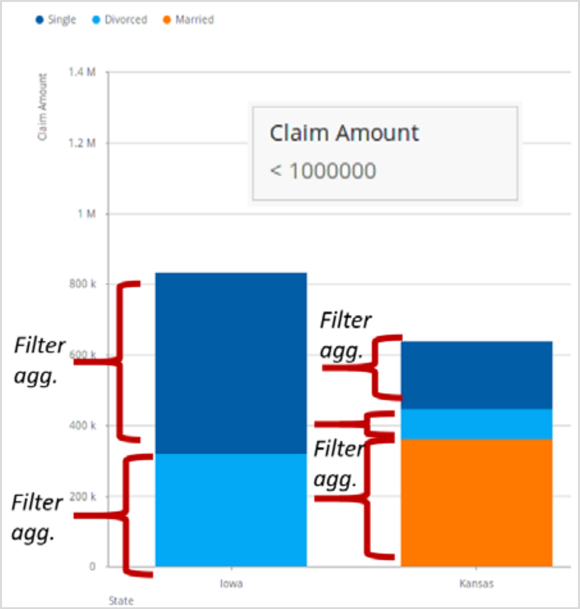
Figure 4: Excluding Null Values

© Hitachi Vantara Corporation 2018. All Rights Reserved

Filtering Aggregations in Tables or Charts

When you apply a filter to a flat table, the filter is applied at the row level.

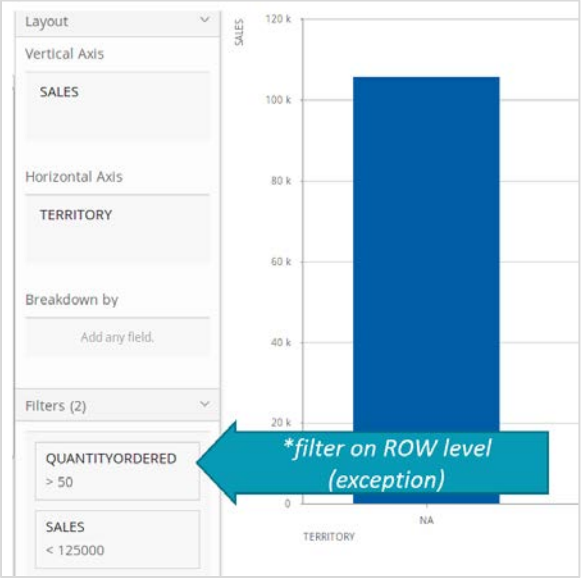
When you apply a filter to charted data, the filter affects each piece of the stack for a bar chart:



Claim Amount
< 1000000

Figure 5: Filtered Bar Chart

For Stream View charts only, filters on numeric fields that are *not* already present on the chart result in the filter being applied at the row level, rather than on the segments of the chart:



Layout

Vertical Axis: SALES

Horizontal Axis: TERRITORY

Breakdown by: Add any field.

Filters (2)

QUANTITYORDERED
> 50

SALES
< 125000

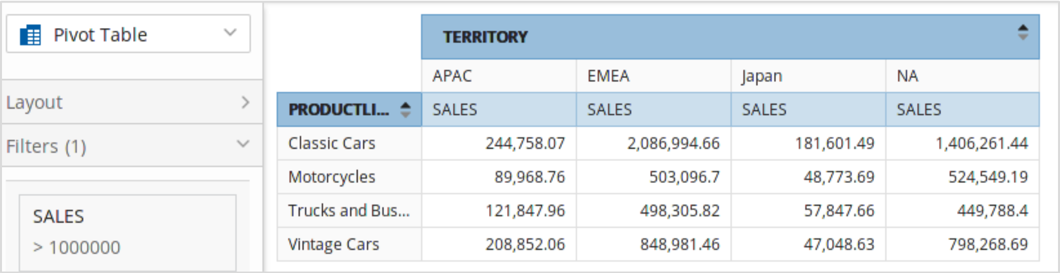
*filter on ROW level (exception)

Figure 6: Filtered Stream View Chart

In pivot tables, numeric filters are first applied to the full aggregation (sum, average...) on the horizontal axis. In the following example, a filter of > `$1 million` has been applied, but only on the

© Hitachi Vantara Corporation 2018. All Rights Reserved

sum/aggregate on the horizontal axis (`Territory = APAC, EMEA`…). This allows subtotals of < `$1 million` to show per `product line`.



SALES
> 1000000

	TERRITORY			
	APAC	EMEA	Japan	NA
PRODUCTLI...	SALES	SALES	SALES	SALES
Classic Cars	244,758.07	2,086,994.66	181,601.49	1,406,261.44
Motorcycles	89,968.76	503,096.7	48,773.69	524,549.19
Trucks and Bus...	121,847.96	498,305.82	57,847.66	449,788.4
Vintage Cars	208,852.06	848,981.46	47,048.63	798,268.69

Figure 7: Filtered Pivot Table

When you apply filters, remember:

- The result sets will change as filter aggregations change.
- Switching from table to chart view keeps already-applied filters.
- Adding a new field on the chart refreshes the filter and alters the results.
- For non-numeric fields, the **contains** and **equals** filters are case-sensitive.

Although we do not support using Pentaho 7.x to open an 8.0 transformation (`.ktr`) that includes DE content, we do support using Pentaho 8.0 to open a 7.x `.ktr` with DE content.

Filtering Limitations

The following limitations exist with filtering in Data Explorer:

Table 1: Filtering Limitations

Limitation	Details
`Date` data type in PDI	It is not possible to apply filters like `equals` or `greater than` to dates using Data Explorer. You can only keep or exclude null values when you are working with dates.
Picking from prepopulated lists of values when filtering on non-numeric fields	This behavior is not supported.
Multi-select	For non-numeric fields, multi-select is only possible with chart actions. When you multi-select (Ctrl + click), you can only select elements from the same area. For example, you can select only labels or only bars in a chart, but not both together.
Row count	The row count in the upper right of Data Explorer does not change as you filter data in and out of view.

© Hitachi Vantara Corporation 2018. All Rights Reserved

Related Information

Here are some links to information that you may find helpful while using this best practices document:

- Inspect Your Data
- Pentaho Analyzer
- Pentaho Components Reference
- Using the Annotate Stream Step for SDR

Finalization Checklist

This checklist is designed to be added to any implemented project that uses this collection of best practices, to verify that all items have been considered and reviews have been performed.

Name of the Project:____

Date of the Review:____

Name of the Reviewer:____

Item	Response	Comments
Did you use Annotate Stream to improve your model results?	YES_______ NO_______	
Did you use filters on your tables or charts?	YES_______ NO_______	

© Hitachi Vantara Corporation 2018. All Rights Reserved