HITACHI
Inspire the Next

Pentaho Data Integration

Data Enrichment

James O'Reilly
Hitachi Vantara Global Learning
Date

© Hitachi Vantara Corporation 2019. All Rights Reserved.

Module Objectives

HITACHI
Inspire the Next

When you complete this module, you should be able to:

- Understand how to implement the various Joins:
  - Cross Join
  - Merge Join
  - Database Join
- Understand the difference between Joins and Lookups
- Enrich the data with scripting steps

© Hitachi Vantara Corporation 2019. All Rights Reserved.

HITACHI
Inspire the Next

Joins

© Hitachi Vantara Corporation 2019. All Rights Reserved.

Topics

HITACHI
Inspire the Next

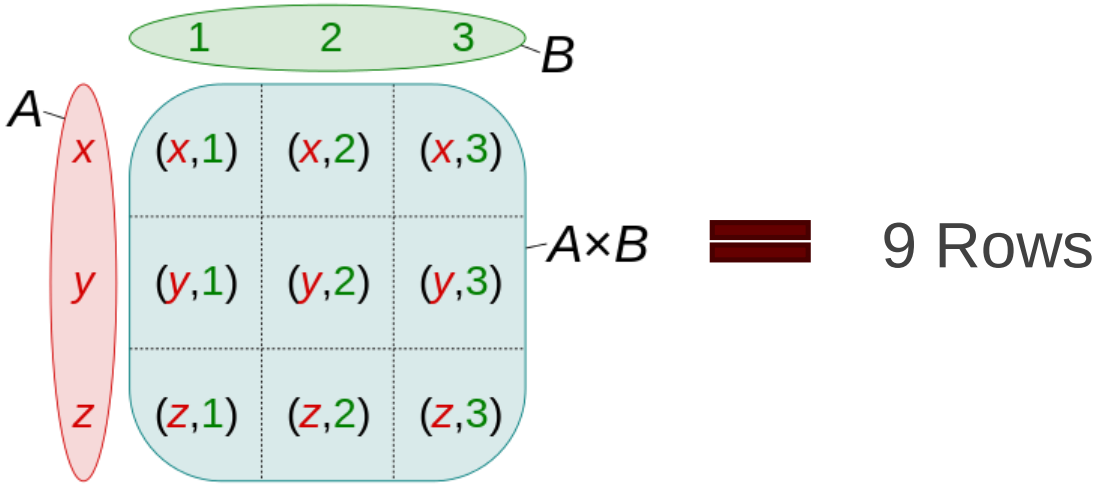
- Merge
- Joins
- Lookups
- Scripting

© Hitachi Vantara Corporation 2019. All Rights Reserved.

Join Rows (Cross | Cartesian)

HITACHI
Inspire the Next

- The Join rows step allows you to produce combinations (Cartesian product) of all rows in the input streams. This normally happens when no matching **join** columns are specified. For example, if table A with 10 rows is joined with table B with 10 rows, a (**Cross) Join** will return 100 rows (Cartesian Product).
  - An alternative with a higher performance in most cases is the Merge Join step.



© Hitachi Vantara Corporation 2019. All Rights Reserved.

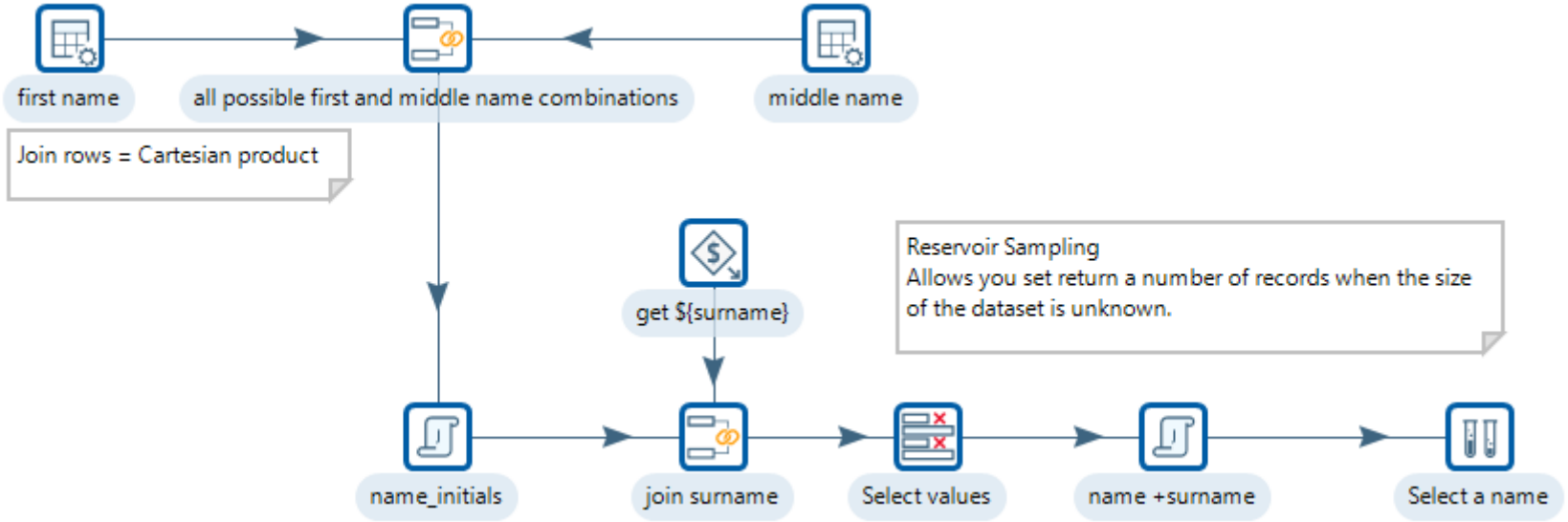
HITACHI
Inspire the Next

Lab 1 – Join Rows

© Hitachi Vantara Corporation 2019. All Rights Reserved.

Lab 1 - Join Rows (Name Selector)

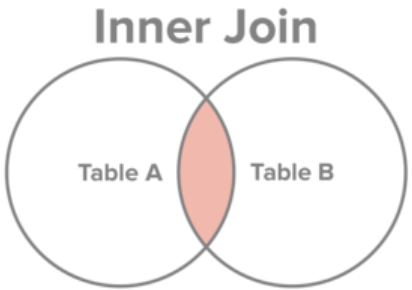
HITACHI
Inspire the Next



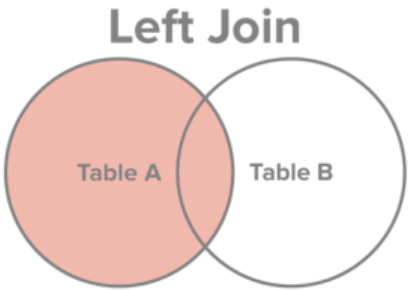
© Hitachi Vantara Corporation 2019. All Rights Reserved.

Examples of SQL Joins

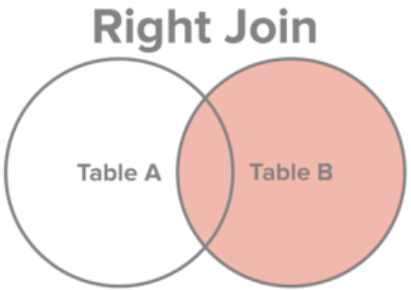
HITACHI
Inspire the Next



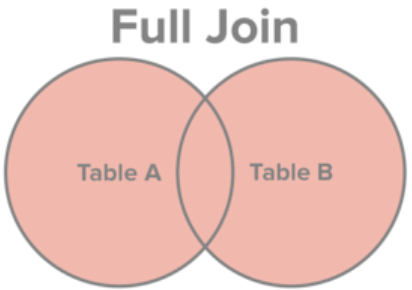
Select all records from Table A and Table B, where the join condition is met.



Select all records from Table A, along with records from Table B for which the join condition is met (if at all).



Select all records from Table B, along with records from Table A for which the join condition is met (if at all).



Select all records from Table A and Table B, regardless of whether the join condition is met or not.

© Hitachi Vantara Corporation 2019. All Rights Reserved.

HITACHI
Inspire the Next

Lab 2 – Join Rows

© Hitachi Vantara Corporation 2019. All Rights Reserved.

Lab 2 - Merge Join

HITACHI
Inspire the Next

- Merge join step performs a merge join between data sets using data from two different input steps.
  - Join options: INNER, LEFT OUTER, RIGHT OUTER, and FULL OUTER

© Hitachi Vantara Corporation 2019. All Rights Reserved.

Merge Join refresher.

HITACHI
Inspire the Next

Table a

ID	Value
1	A
2	B
3	C
4	D

Table b

ID	Value
1	Red
3	Blue
5	Yellow

a	b
A	Red
C	Blue

A regular INNER join between two tables will produce a result set with only the common values from both tables.

a	b
A	Red
B	NULL
C	Blue
D	NULL

A LEFT OUTER join will display all the values from the left table, matching values from the right table, and inserting NULL values for non-matching values.

© Hitachi Vantara Corporation 2019. All Rights Reserved.

Merge Join refresher.

HITACHI
Inspire the Next

Table a

ID	Value
1	A
2	B
3	C
4	D

Table b

ID	Value
1	Red
3	Blue
5	Yellow

a	b
A	Red
C	Blue
NULL	Yellow

A RIGHT OUTER join will display all the values from the right table, matching values from the left table, and inserting NULL values for non-matching values.

a	b
A	Red
B	NULL
C	Blue
D	NULL
NULL	Yellow

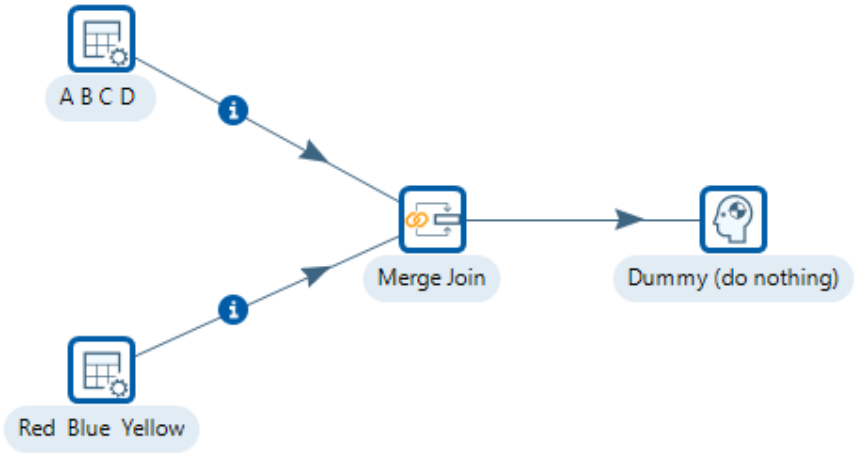
A full OUTER join is not available in MySQL. It takes the values from both tables, inserting NULL values for non-matching values.

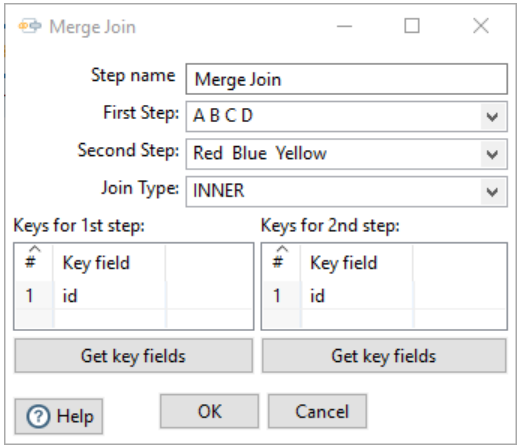
© Hitachi Vantara Corporation 2019. All Rights Reserved.

Lab 2 - Merge Join - Overview

HITACHI
Inspire the Next

Simple merge join demonstration.



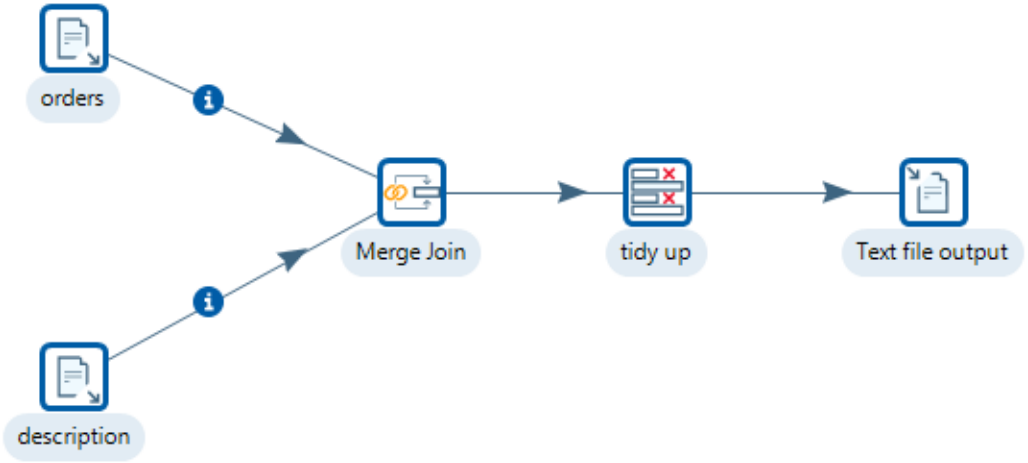


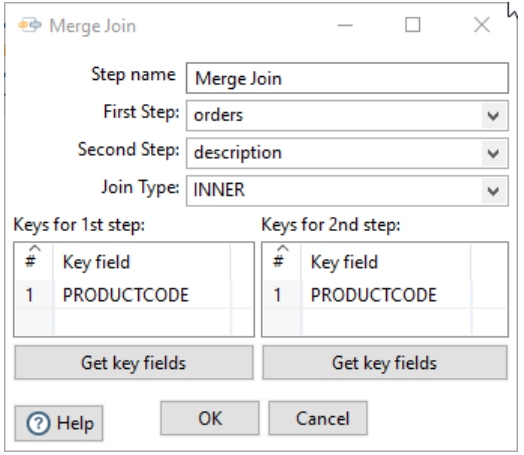
© Hitachi Vantara Corporation 2019. All Rights Reserved.

Lab 2 - Merge Join - Orders

HITACHI
Inspire the Next

Solves the problem of merge rows.. GD4-1-1
With a 'join' you dont have to ensure that each data stream has the same structure / layout.





© Hitachi Vantara Corporation 2019. All Rights Reserved.

HITACHI
Inspire the Next

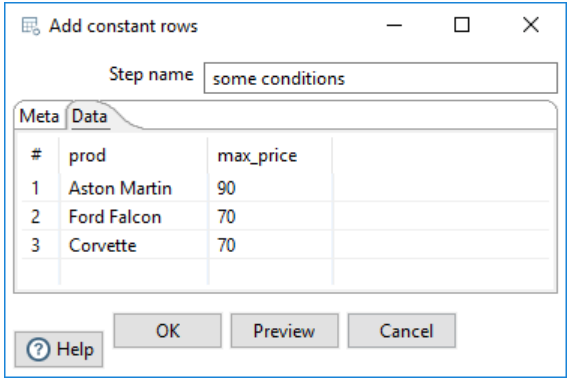
Lab 3 – Database Join

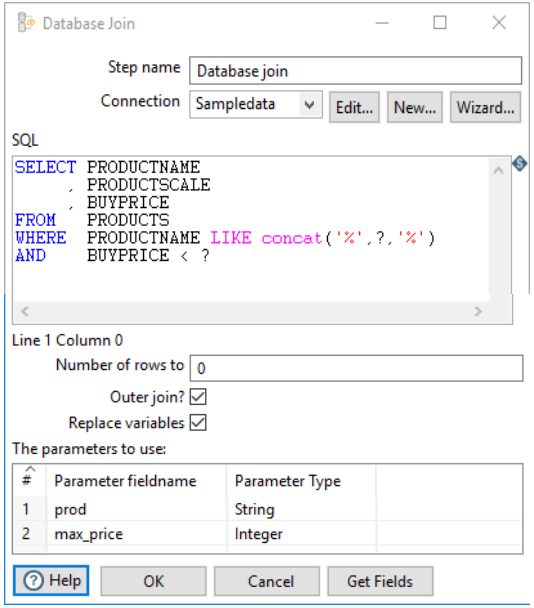
© Hitachi Vantara Corporation 2019. All Rights Reserved.

Lab 3 - Database Join

HITACHI
Inspire the Next

- Searching for information in databases, text files, web services, and so on, is a very common task.
- The Database Join compares a dataset of records against the table using variables / parameters.





As its uses an Outer Join, All the rows from the datasets are returned.

© Hitachi Vantara Corporation 2019. All Rights Reserved.

Lab 3 - Database Join

HITACHI
Inspire the Next

- The first record

```
WHERE PRODUCTNAME LIKE concat ('%','Aston Martin','%') AND BUYPRICE < 90
```

The other database fields are returned as stream fields.

- The second record

```
WHERE PRODUCTNAME LIKE concat ('%',‘Ford Falcon','%') AND BUYPRICE < 70
```

No ‘Ford Falcon’ values were found with a BUYPRICE <70, so the fields returned NULL values.

- The third record

```
WHERE PRODUCTNAME LIKE concat ('%’,‘Chevette','%') AND BUYPRICE < 70
```

The other database fields are returned as stream fields.

© Hitachi Vantara Corporation 2019. All Rights Reserved.

HITACHI
Inspire the Next

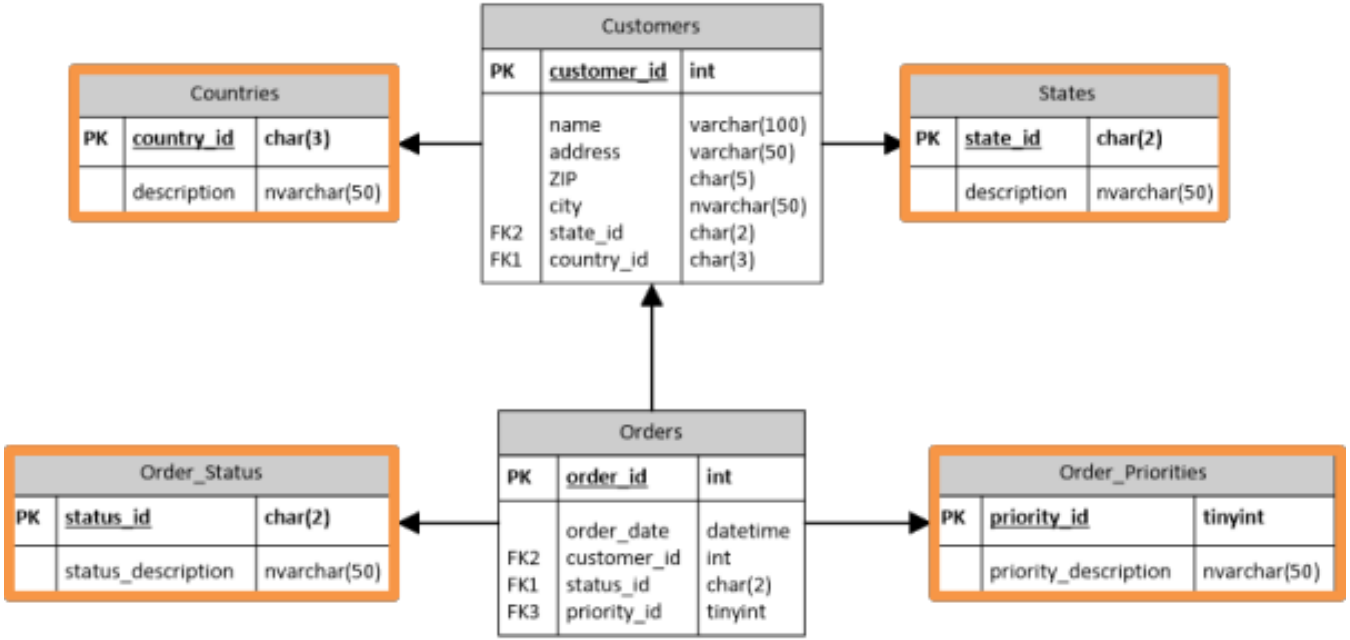
Lookups

© Hitachi Vantara Corporation 2019. All Rights Reserved.

Database Lookup Tables

HITACHI
Inspire the Next

- Besides transforming the data, you may need to search and bring data from other sources.



© Hitachi Vantara Corporation 2019. All Rights Reserved.

HITACHI
Inspire the Next

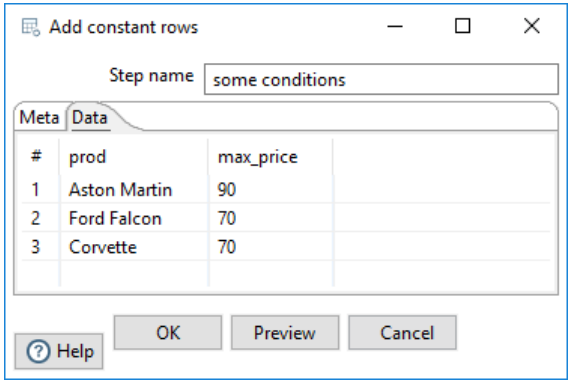
Lab 1 – Database Lookups

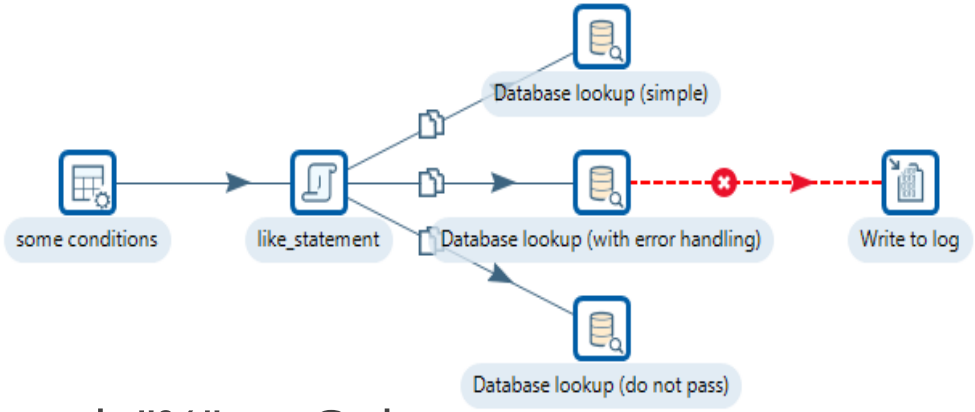
© Hitachi Vantara Corporation 2019. All Rights Reserved.

Lab 1 - Database Lookup

HITACHI
Inspire the Next

- Simple - straightforward lookup, returns a single
- Error - rows that don't compare are streamed to error step
- Do not pass – don't pass error





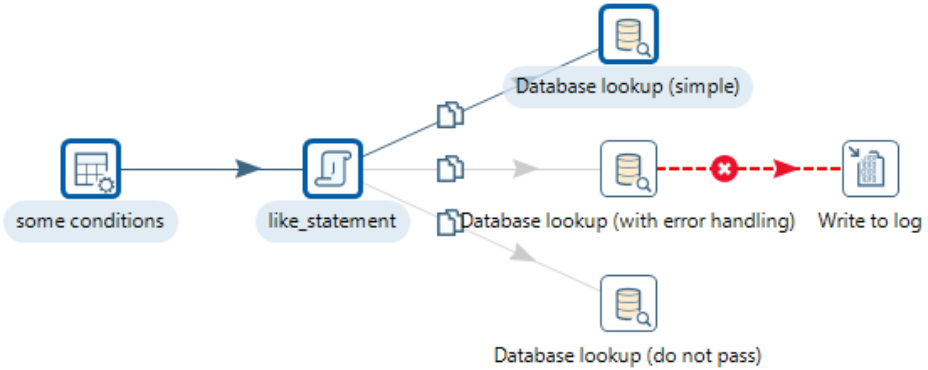
- like_statement "%"+prod+"%" String

© Hitachi Vantara Corporation 2019. All Rights Reserved.

Lab 1 - Database Lookup

HITACHI
Inspire the Next

- Simple



Execution Results

Execution History | Logging | Step Metrics | Performance Graph | Metrics | Preview data

First rows | Last rows | Off — Inspect Data

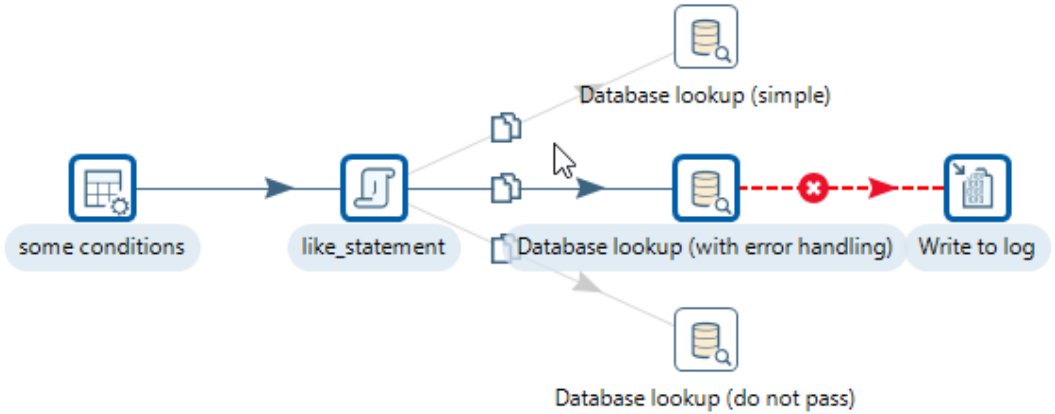
#	prod	max_price	like_statement	PRODUCTNAME	PRODUCTSCALE
1	Aston Martin	90	%Aston Martin%	1965 Aston Martin DB5	1:18
2	Ford Falcon	70	%Ford Falcon%	not available	<null>
3	Corvette	70	%Corvette%	1958 Chevy Corvette Limited Edition	1:24

© Hitachi Vantara Corporation 2019. All Rights Reserved.

Lab 1 - Database Lookup

HITACHI
Inspire the Next

- Error




```
2016/12/10 18:26:48 - Spoon - Launching transformation [GD4-2-1_databse_lookup]...
2016/12/10 18:26:48 - Spoon - Started the transformation execution.
2016/12/10 18:26:48 - GD4-2-1_databse_lookup - Dispatching started for transformation [GD4-2-1_databse_lookup]
2016/12/10 18:26:48 - some conditions.0 - Finished processing (I=0, O=0, R=0, W=3, U=0, E=0)
2016/12/10 18:26:48 - like_statement.0 - Finished processing (I=0, O=0, R=3, W=3, U=0, E=0)
2016/12/10 18:26:48 - Write to log.0 -
2016/12/10 18:26:48 - Write to log.0 - ------------> Linenr 1----------------------------
2016/12/10 18:26:48 - Write to log.0 - no productname match at that price
2016/12/10 18:26:48 - Write to log.0 -
2016/12/10 18:26:48 - Write to log.0 - Ford Falcon
2016/12/10 18:26:48 - Write to log.0 - 70
2016/12/10 18:26:48 - Write to log.0 - %Ford Falcon%
2016/12/10 18:26:48 - Write to log.0 -
2016/12/10 18:26:48 - Write to log.0 - ====================
2016/12/10 18:26:48 - Database lookup (with error handling).0 - Finished processing (I=2, O=0, R=3, W=2, U=0, E=1)
2016/12/10 18:26:48 - Write to log.0 - Finished processing (I=0, O=0, R=1, W=1, U=0, E=0)
2016/12/10 18:26:48 - Spoon - The transformation has finished!!
```

Execution Results

Execution History | Logging | Step Metrics | Performance Graph | Metrics | Preview data

First rows | Last rows | Off — Inspect Data

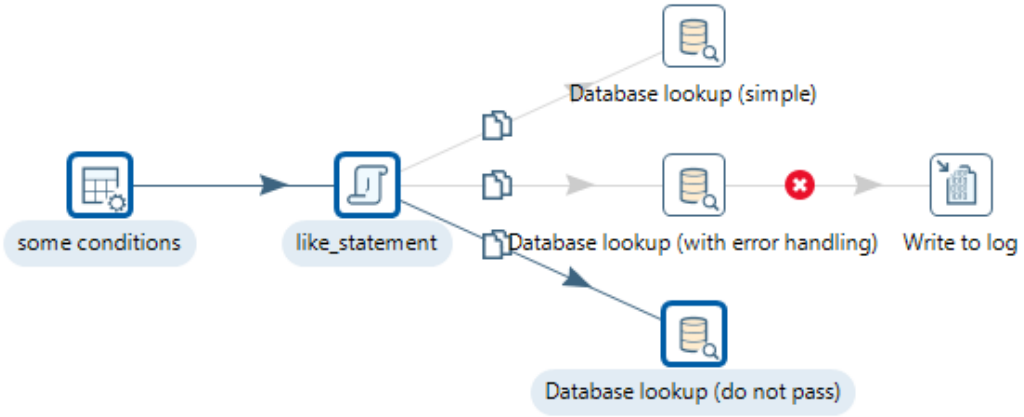
#	prod	max_price	like_statement	PRODUCTNAME	PRODUCTSCALE
1	Aston Martin	90	%Aston Martin%	1965 Aston Martin DB5	1:18
2	Corvette	70	%Corvette%	1958 Chevy Corvette Limited Edition	1:24

© Hitachi Vantara Corporation 2019. All Rights Reserved.

Lab 1 - Database Lookup

HITACHI
Inspire the Next

- Do not pass



Execution Results

Execution History | Logging | Step Metrics | Performance Graph | Metrics | Preview data

First rows | Last rows | Off | Inspect Data

#	prod	max_price	like_statement	PRODUCTNAME	PRODUCTSCALE
1	Aston Martin	90	%Aston Martin%	1965 Aston Martin DB5	1:18
2	Corvette	70	%Corvette%	1958 Chevy Corvette Limited Edition	1:24

© Hitachi Vantara Corporation 2019. All Rights Reserved.

HITACHI
Inspire the Next

Scripting

© Hitachi Vantara Corporation 2019. All Rights Reserved.

Scripting

HITACHI
Inspire the Next

- Love-hate relationship: maintainability vs. power and flexibility
- Historically, Java Script step was PDI's "duct tape", taking care of complex transformation work
- Over the years more standard steps and job entries got introduced
- As a general rule of thumb, avoid using scripting altogether

© Hitachi Vantara Corporation 2019. All Rights Reserved.

Lab 1 – Formula

© Hitachi Vantara Corporation 2019. All Rights Reserved.

Lab 1 - Formula Step

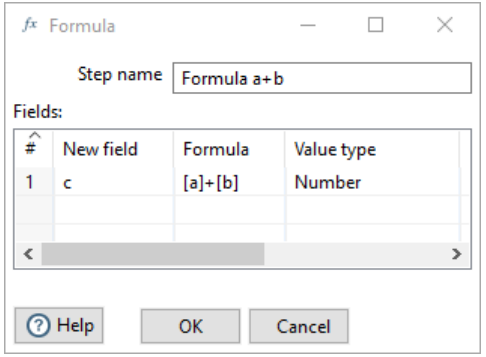
HITACHI
Inspire the Next

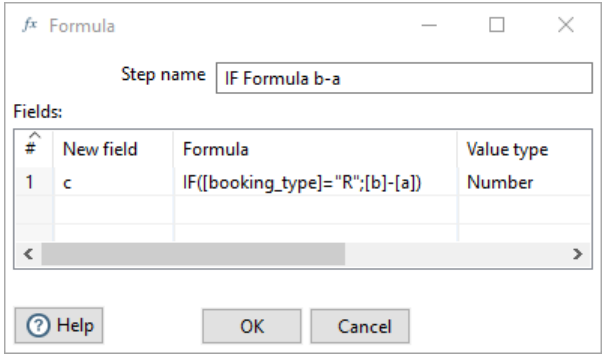
- Oasis OpenFormula syntax (also used in Calc, OpenOffice)

  http://docs.oasis-open.org/office/v1.2/OpenDocument-v1.2-part2.html

- Allows for more flexible formulas than the predefined ones in the Calculator step

- Conditional logic
  - No fields selector

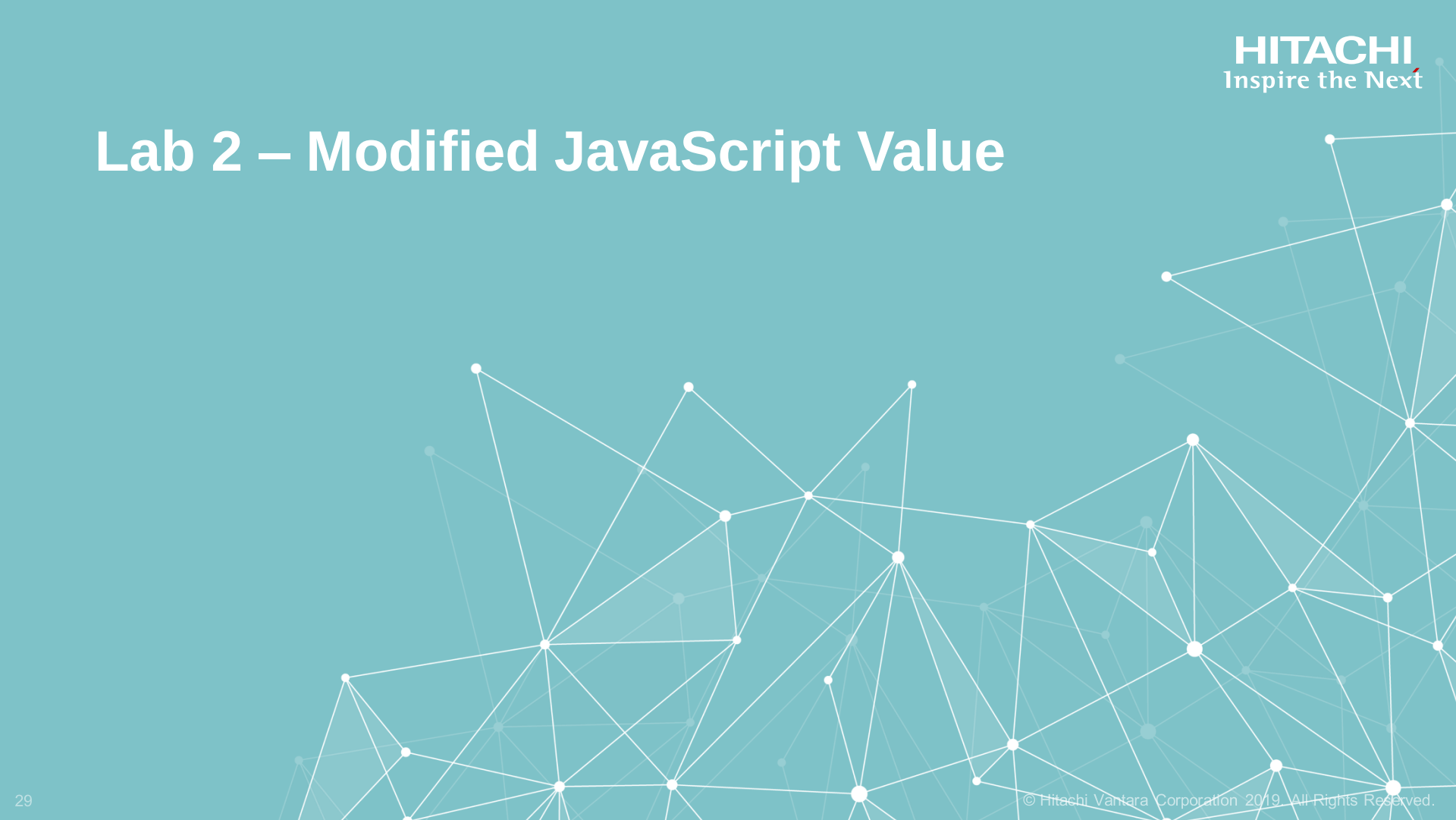




© Hitachi Vantara Corporation 2019. All Rights Reserved.

HITACHI
Inspire the Next

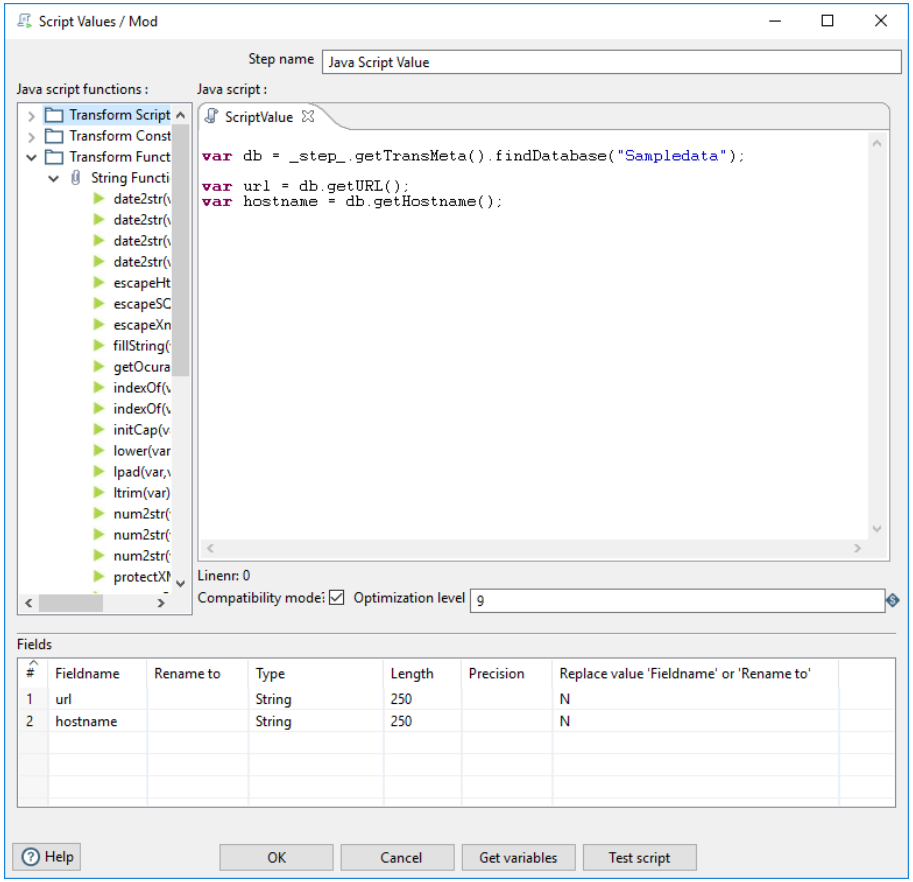
Lab 2 – Modified JavaScript Value



© Hitachi Vantara Corporation 2019. All Rights Reserved.

Modified JavaScript Value

- Using JavaScript in a transformation.
- Lots of inbuilt functions..
- Look in the samples folder ..



© Hitachi Vantara Corporation 2019. All Rights Reserved.

Compatibility Mode

HITACHI
Inspire the Next

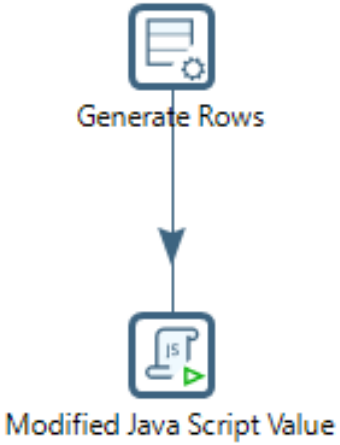
What does the compatibility switch do?

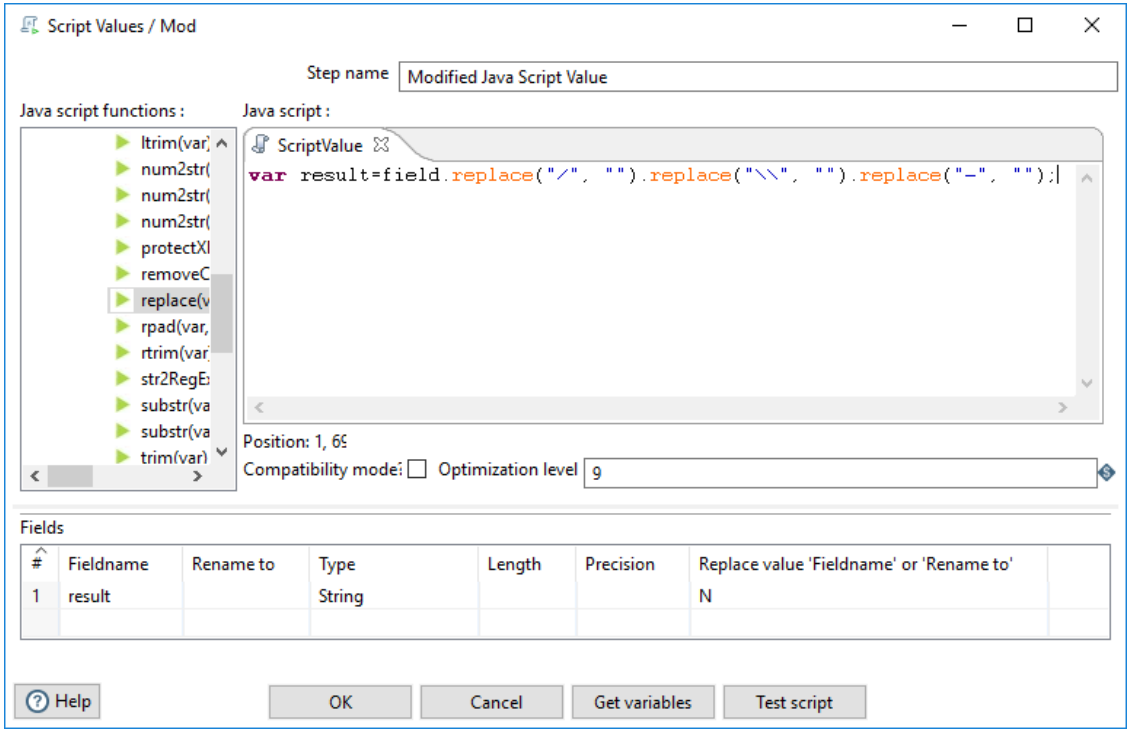
- There are two version of the JavaScript engine: the 2.5 version and the 3 version. If "compatibility mode" is checked (and by default it is), JavaScript works like it did in version 2.5. Obviously the new version should be used if possible so uncheck "compatibility mode" if you can.

- The big difference between the two versions is that in 2.5, value objects are directly modifiable and their type can be changed (a date variable can be converted into a string). This can cause errors. Because this is no longer possible in the 3.0 version, the JavaScript should also be faster.

© Hitachi Vantara Corporation 2019. All Rights Reserved.

Lab 3 - Modified JavaScript Value





© Hitachi Vantara Corporation 2019. All Rights Reserved.

HITACHI
Inspire the Next

Lab 3 – Java Class

© Hitachi Vantara Corporation 2019. All Rights Reserved.

User Defined Java Class

HITACHI
Inspire the Next

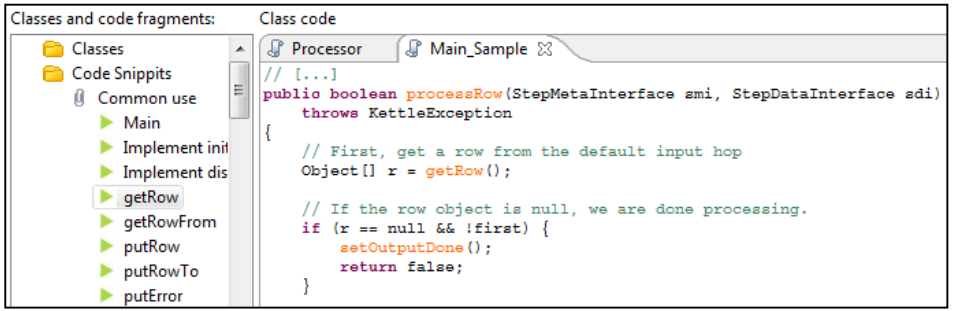
- Instead of just a single expression, this step lets you define a complete class, which allows you to write a Kettle plugin as a step.

  http://rpbouman.blogspot.co.uk/2009/11/pentaho-data-integration-javascript.html

  http://wiki.pentaho.com/display/EAI/Writing+your+own+Pentaho+Data+Integration+Plug-In

- The benefit of User Defined Java Class is to simplify the deployment process.

- Code snippets provide samples:

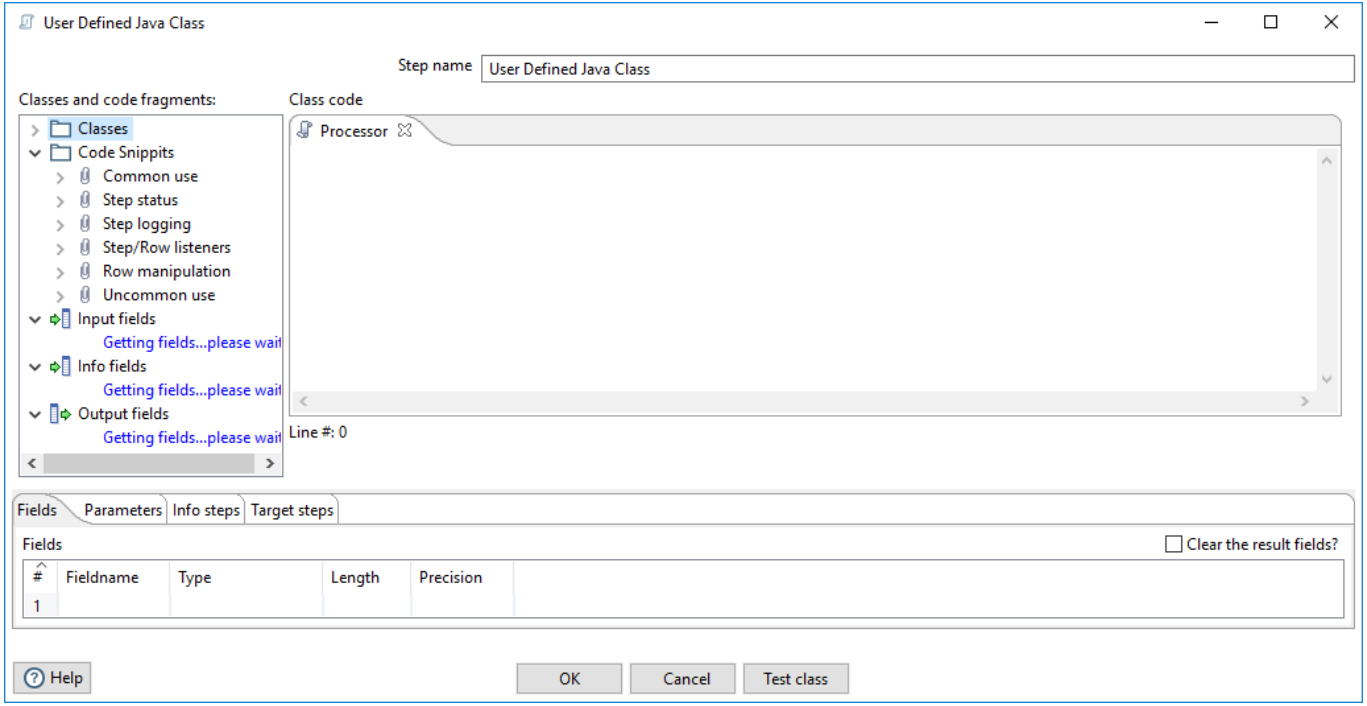


© Hitachi Vantara Corporation 2019. All Rights Reserved.

User Defined Java Class

HITACHI
Inspire the Next

- Code snippet samples



© Hitachi Vantara Corporation 2019. All Rights Reserved.

Module Recap

HITACHI
Inspire the Next

- In this module, you should have learned to:
- Understand how to implement the various Joins:
  - Cross Join
  - Merge Join
  - Database Join
- Understand the difference between Joins and Lookups
- Enrich the data with scripting steps

© Hitachi Vantara Corporation 2019. All Rights Reserved.

HITACHI
Inspire the Next

Thank You

© Hitachi Vantara Corporation 2019. All Rights Reserved.

HITACHI

Inspire the Next