HITACHI
Inspire the Next

Pentaho Data Integration

Working with Flat Files

James O'Reilly
Hitachi Vantara Global Learning
Date

Module Objectives

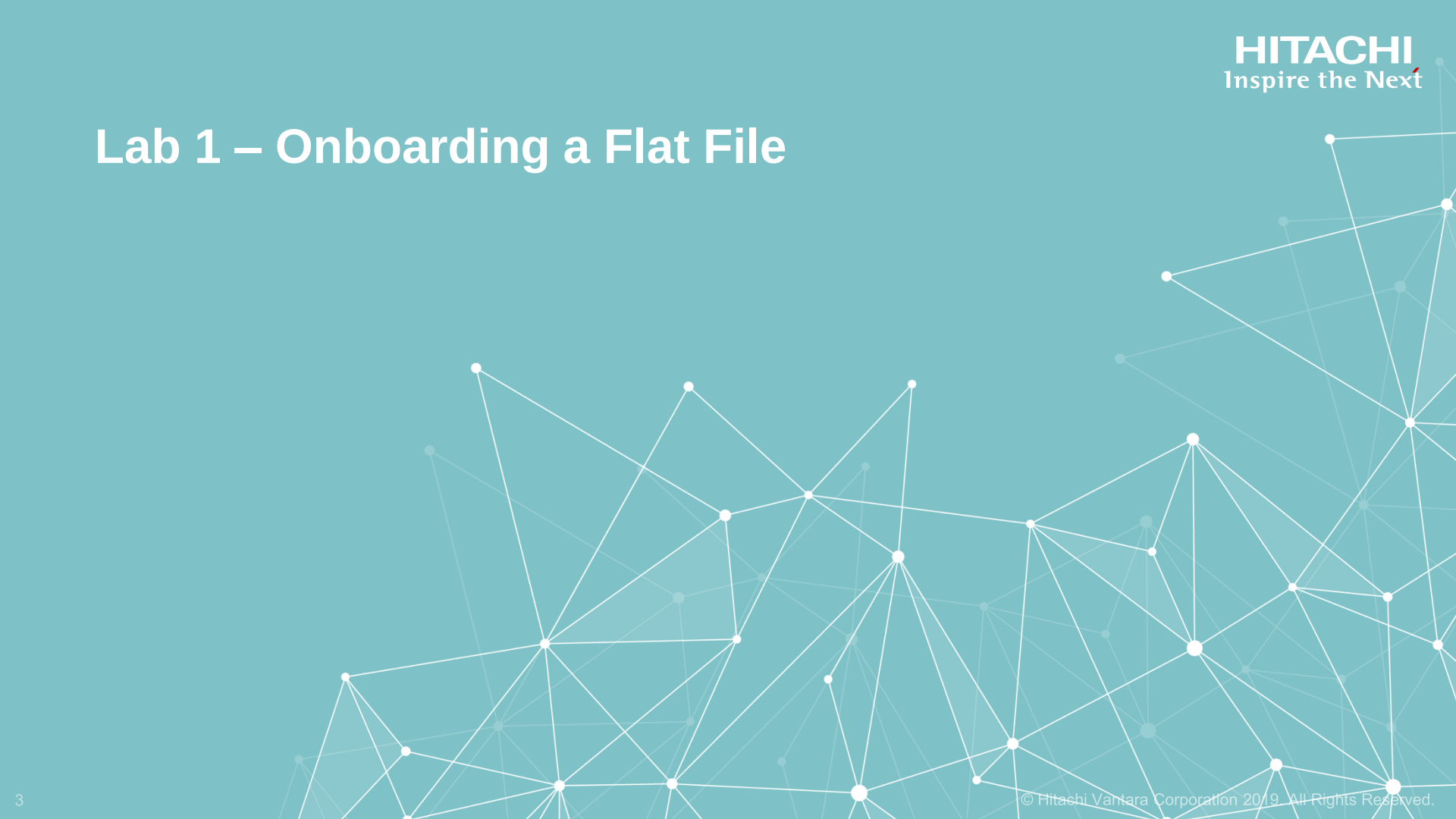
HITACHI
Inspire the Next

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

- Configure steps to onboard various file formats:
  - Onboard a TXT / CSV
  - Output a TXT / CSV file
  - Create Excel Workbooks based on templates
  - Onboard XML & JSON as datasources

HITACHI
Inspire the Next

Lab 1 – Onboarding a Flat File



Lab 1: Text File / CSV Input

HITACHI
Inspire the Next

- Create an ETL workflow that will write the data to a database table:

```
Productline: Classic Cars
Customer: Christine Loomis
Delivered: January 2004
Order Value: $21.99

Productline: Classic Cars
Customer: Mary L. Peachin
Delivered: November 2008
Order Value: $24.99

Productline: Trains
Customer: Bob Italia
Delivered: July 1994
Order Value: $14.99

Productline: Planes
Customer: Scott M. Ascher
Delivered:  March 2014
Order Value: $27.99

Productline: Motorcycles
Customer: Monty Halls
Returned: April 2007
Order Value: $29.99

Productline: Trains
Customer: Paul McCallum
Returned: June 2017
Order Value: $34.99

Productline: Boats
Customer: Jill Robinson
Delivered: November 2014
Order Value: $19.99
```

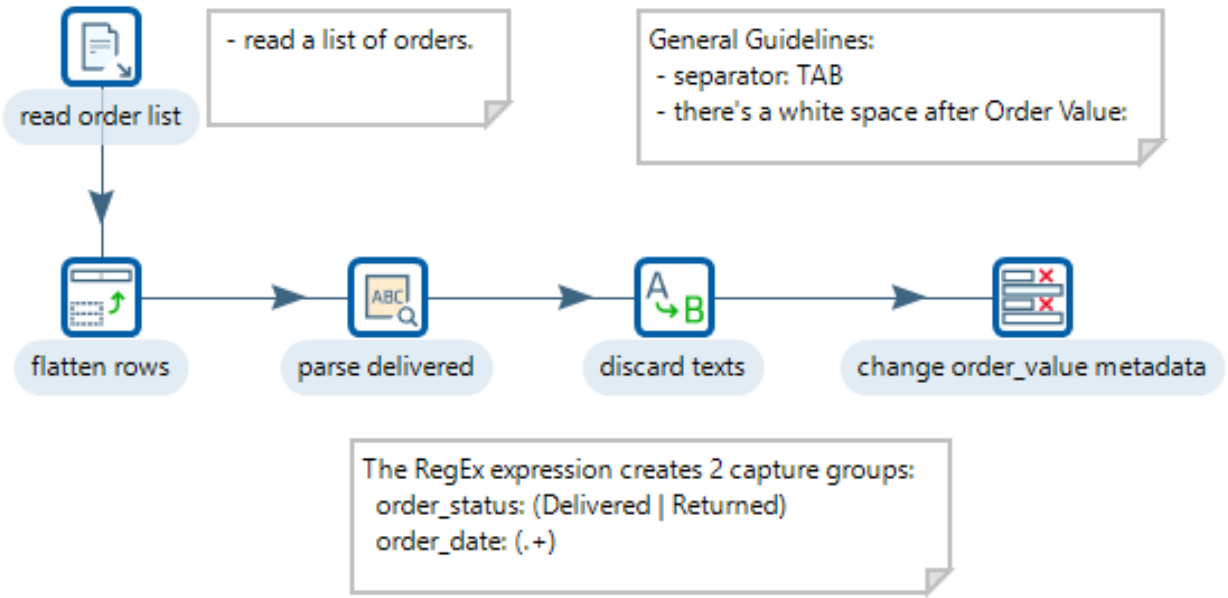
So what approach would you recommend?

Flatten the layout for each data stream column:
Productline Customer Status Order_Value Order_Date

Status can have a value of either: Delivered | Returned

Lab 1: Text File / CSV Input

- Onboard data
- Flatten Rows
- Capture Groups
- Trim Text
- Select Values



HITACHI
Inspire the Next

Lab 2 – Creating a Flat File

Lab 2: Write to a Text File

HITACHI
Inspire the Next

- Create a survey based on questions in a text file.

```
text
---------------------------------------------------------
Please answer the questions below:
```

Head

Stream Append

```
How many employees currently in your organisation?
Which ETL tool do you currently use?
What would you do if you had a magic wand?
```

Body

So what approach would you take?

In the ‘head’ workflow you have the input for the Customer Name.
In the ‘body’ workflow you have the questions.

Then append the ‘header’ stream to the ‘body’ stream

Lab 2: Write to a Text File

HITACHI
Inspire the Next

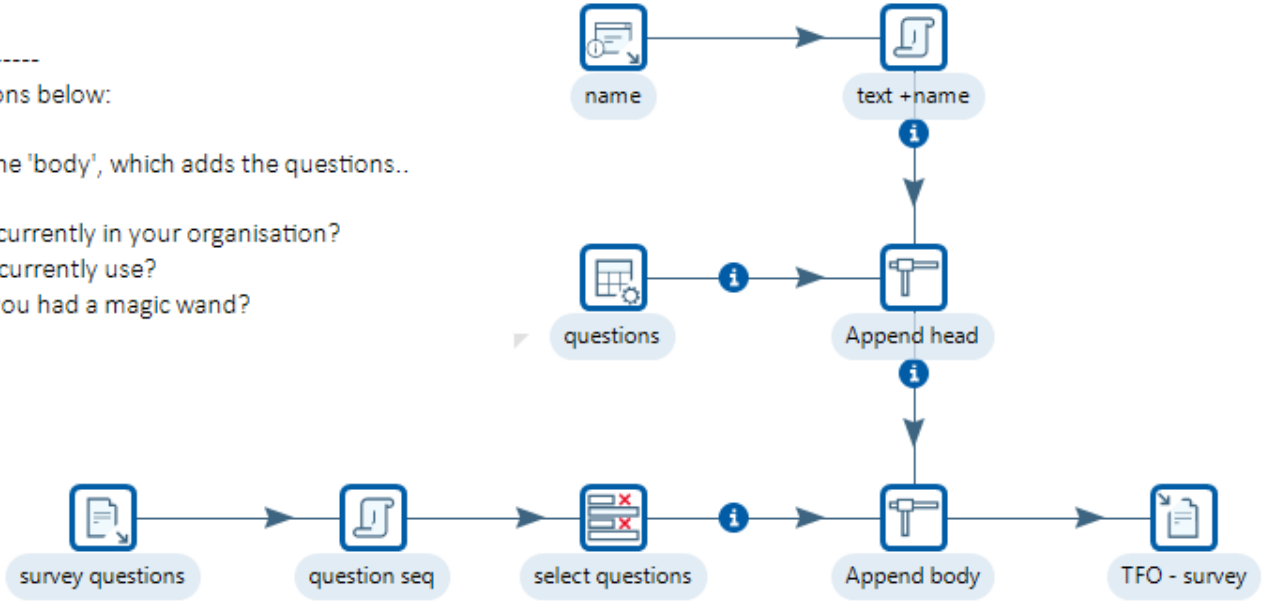
This guided demonstration illustrates how to write an unstructured file.
The first part of the transformation, defines the 'head':

Customer name:

Please answer the questions below:

Which is then apended to the 'body', which adds the questions..

1. How many employees currently in your organisation?
2. Which ETL tool do you currently use?
3. What would you do if you had a magic wand?



HITACHI
Inspire the Next

Lab 3 – Excel

Lab 3: Excel

HITACHI
Inspire the Next

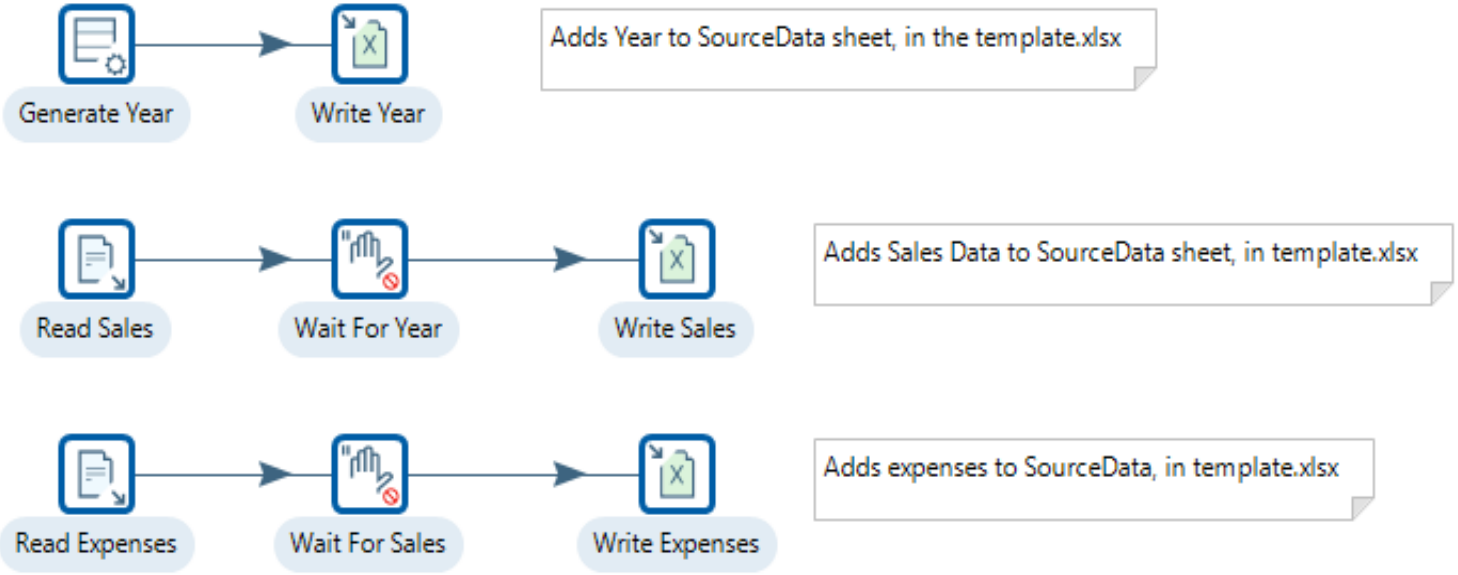
- Steel Wheels wish to automate their Half Yearly Sales and Expenses Report (Excel).
  - Sales and Expenses are text files.
  - Leverage a template.

	A	B	C	D	E	F	G	H	I	J	K
1	Year	2016									
2											
3		JANUARY	FEBRUARY	MARCH	APRIL	MAY	JUNE	6-MONTH TOTAL	MEAN	MINIMUM	MAXIMUM
4	PRODUCTLINE										
5	Classic Cars	23,455.22€	25,442.11€	24,222.89€	20,233.11€	19,876.22€	19,233.56€	132,463.11€	22,077.19€	19,233.56€	25,442.11€
6	Motorcycles	11,244.21€	12,987.69€	13,954.09€	13,006.33€	13,065.31€	12,087.74€	76,345.37€	12,724.23€	11,244.21€	13,954.09€
7	Trains	1,231.29€	1,227.98€	1,395.33€	1,399.90€	1,335.90€	1,376.98€	7,967.38€	1,327.90€	1,227.98€	1,399.90€
8	Planes	956.12€	834.56€	457.76€	765.32€	898.11€	667.49€	4,579.36€	763.23€	457.76€	956.12€
9											
10	Total Sales	35,655.55€	39,264.36€	38,634.74€	34,004.76€	33,839.64€	31,988.79€	213,387.84€	35,564.64€	31,988.79€	39,264.36€
11		2,055.22€	2,542.11€	2,422.89€	2,033.11€	1,986.22€	1,933.56€				
12	EXPENSES	100.32€	103.23€	140.23€	130.23€	120.33€	121.34€				
13	Advertising	11,020.80€	11,020.80€	11,020.80€	9,350.10€	9,350.10€	12,350.60€	64,113.20€	10,685.53€	9,350.10€	12,350.60€
14	Cost of Goods	223.23€	223.23€	223.23€	223.23€	223.23€	223.23€	1,339.38€	223.23€	223.23€	223.23€
15	Salary	10.30€	0.00€	209.99€	3.99€	0.00€	12.23€	236.51€	39.42€	0.00€	209.99€
16	Lease	90.23€	90.23€	78.90€	90.23€	78.90€	0.00€	428.49€	71.42€	0.00€	90.23€
17	Miscellaneous	0.00€	0.00€	0.00€	0.00€	0.00€	0.00€	0.00€	0.00€	0.00€	0.00€
18	Overhead	0.00€	0.00€	0.00€	0.00€	0.00€	0.00€	0.00€	0.00€	0.00€	0.00€
19	Total Expenses	11,344.56€	11,334.26€	11,532.92€	9,667.55€	9,652.23€	12,586.06€	66,117.58€	11,019.60€	9,652.23€	12,586.06€
20											
21	PROFIT	24,310.99€	27,930.10€	27,101.82€	24,337.21€	24,187.41€	19,402.73€	147,270.26€	24,545.04€	19,402.73€	27,930.10€

Lab 3: Excel

HITACHI
Inspire the Next

This guided demonstration creates an Excel workbook based on a template that is populated from several Excel spreadsheets, with text files as their datasource.



HITACHI
Inspire the Next

Lab 4 – XML

Lab 4: XML

HITACHI
Inspire the Next

- XML stands for EXtensible Markup Language.
- XML documents are used to not only store data, but exchange data between systems.

```
<?xml version="1.0" encoding="UTF-8"?>
<document>
<order>
  <productline>Classic Cars</productline>
  <customer>Christine Loomis</customer>
  <status>Delivered</status>
  <date>January 2004</date>
  <value>21.99</value>
</order>
```

X-path to attributes

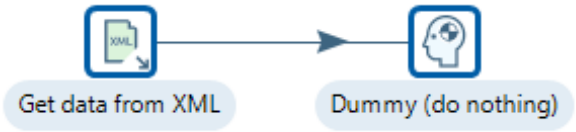
Lab 4: XML

HITACHI
Inspire the Next

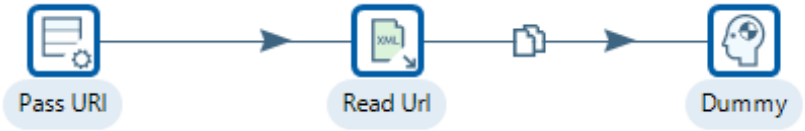
‘Get data From XML’ step can read data from 3 kind of sources:

- file
- url
- stream

A simple example of reading an xml file



Reading xml from a URL



HITACHI
Inspire the Next

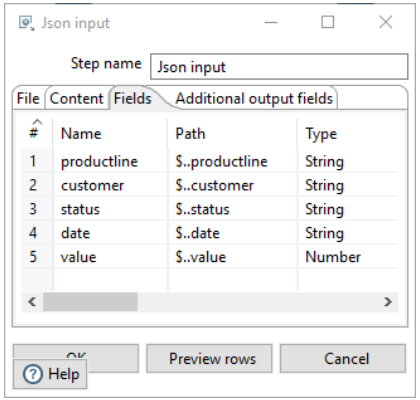
Lab 5 – JSON

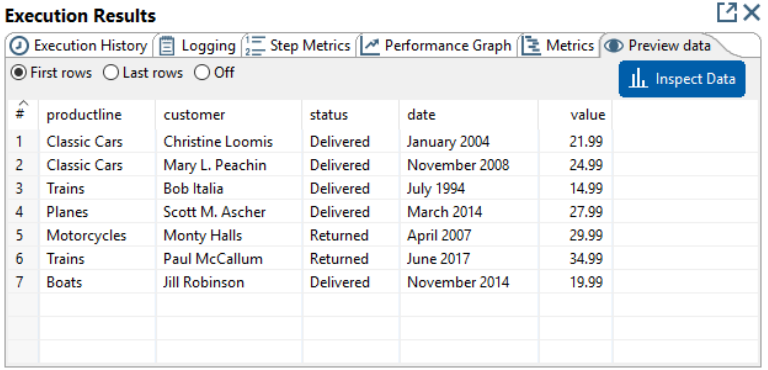
Lab 5: JSON

HITACHI
Inspire the Next

Read Json file and extract portions data out of structure.


```
{ "document": {
    "order": [
      { "productline": "Classic Cars",
        "customer": "Christine Loomis",
        "status": "Delivered",
          "date": "January 2004",
        "value": 21.99
      },
      { "productline": "Classic Cars",
        "customer": "Mary L. Peachin",
        "status": "Delivered",
          "date": "November 2008",
        "value": 24.99
      },
      { "productline": "Trains",
        "customer": "Bob Italia",
        "status": "Delivered",
        "date": "July 1994",
        "value": 14.99
      },
      { "productline": "Planes",
        "customer": "Scott M. Ascher",
        "status": "Delivered",
        "date": "March 2014",
        "value": 27.99
      },
```



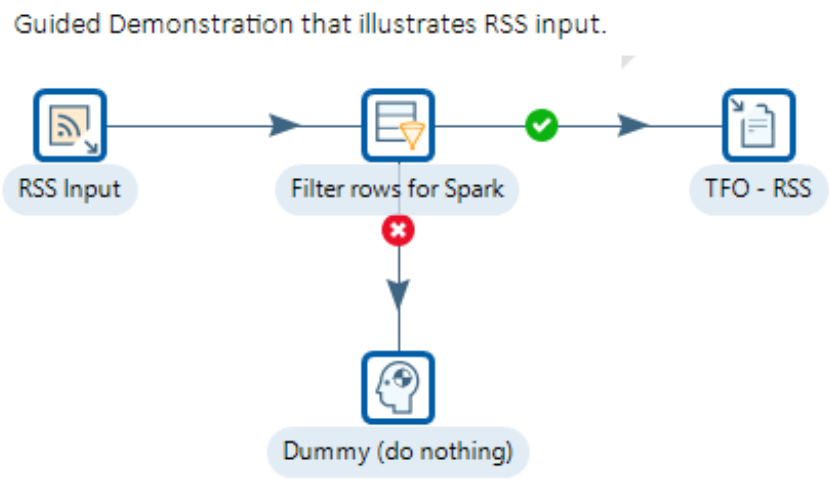
HITACHI
Inspire the Next

Lab 6 – RSS Feed

Lab 6: RSS Feed

HITACHI
Inspire the Next

- RSS (Rich Site Summary) is a format for delivering regularly changing web content. Many news-related sites, weblogs and other online publishers syndicate their content as an RSS Feed to whoever wants it.
  - Filter for ‘Titles that contain XXX’.



Module Objectives

HITACHI
Inspire the Next

In this module, you should have learned to:

- Configure steps to onboard various file formats:
  - Onboard a TXT / CSV
  - Output a TXT / CSV file
  - Create Excel Workbooks based on templates
  - Onboard XML & JSON as data sources

HITACHI
Inspire the Next

Thank You

HITACHI

Inspire the Next