
Updates on webSpoon and other innovations from Hitachi R&D

11/11/2017

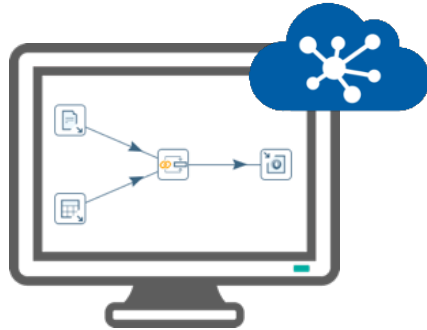
Hiromu Hota, PhD

@HiromuHota, hiromu.hota@hal.hitachi.com
Researcher at Hitachi America, Ltd.

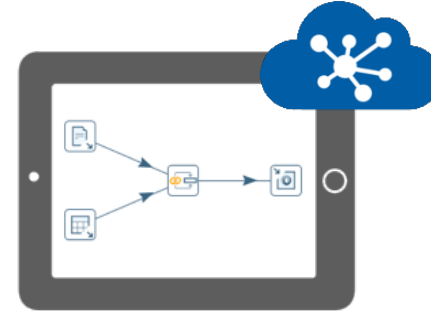
- webSpoon
 - Demo
 - Updates since PCM16 and missings
 - Use cases

webSpoon: a browser-based Spoon

- webSpoon works on any latest browser, accessible over a network.



Desktop/laptop



Smartphone/tablet

- webSpoon has advantages:



✓ Data security



✓ Remote use



✓ Ease of mgmt.






✓ Cloud

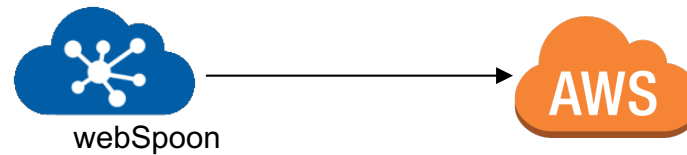
- webSpoon is **NOT** supported by Pentaho or by Hitachi.

Demo

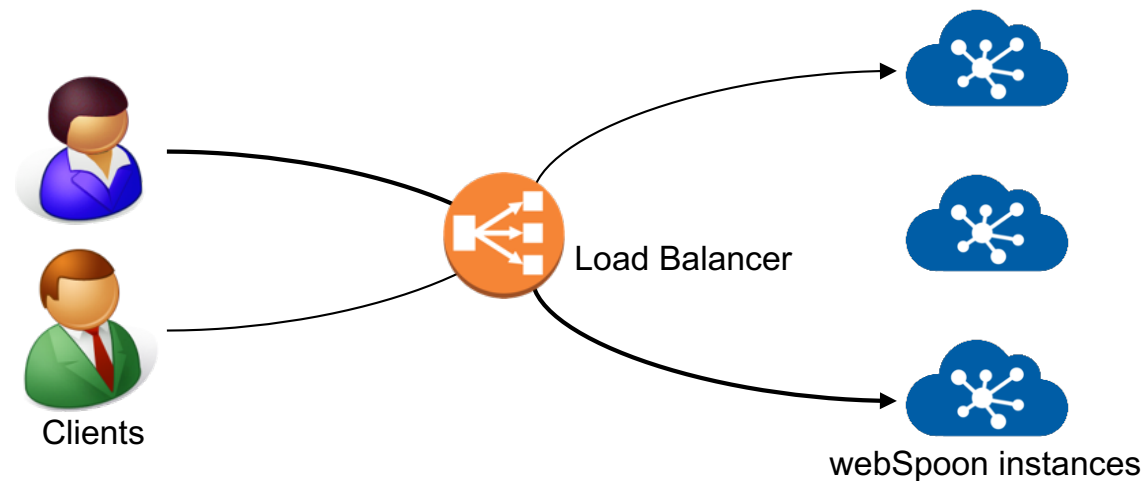
1. webSpoon (demo instance in AWS)
2. Multi-tenancy (demo instance in local Docker)

- webSpoon became matured in its stability, functionality, and usability.
- Stability
 - Fixed many things: menubar, shortcuts, scrollbar/zooming, copy/paste
 - Automated UI Testing
 - CI/CD (nightly build for every commit)  **Jenkins**
- Functionality
 - Lots of steps/job entries and other type of plugins confirmed to be compatible
 - Carte integration
 - Multi-user/-tenant 
- Usability
 - FileDialog to open from the server's file system / Import from the client's
 - Dockerized 
 - No longer 9051 port mapping

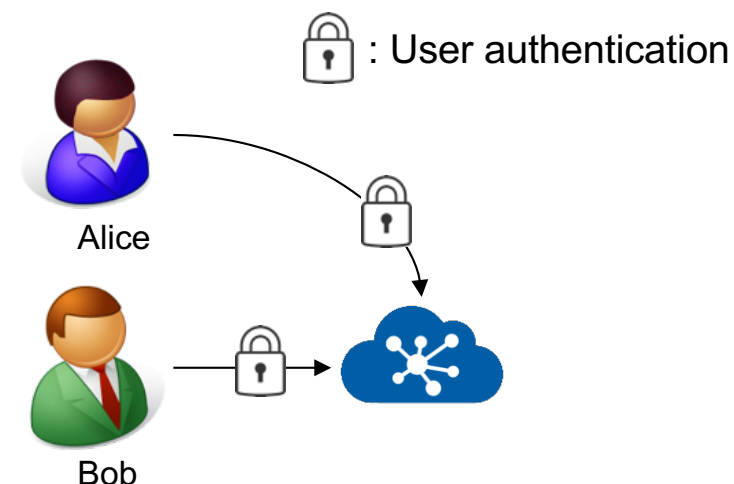
- webSpoon is easily deployable to the cloud
 - E.g., AWS Elastic Beanstalk



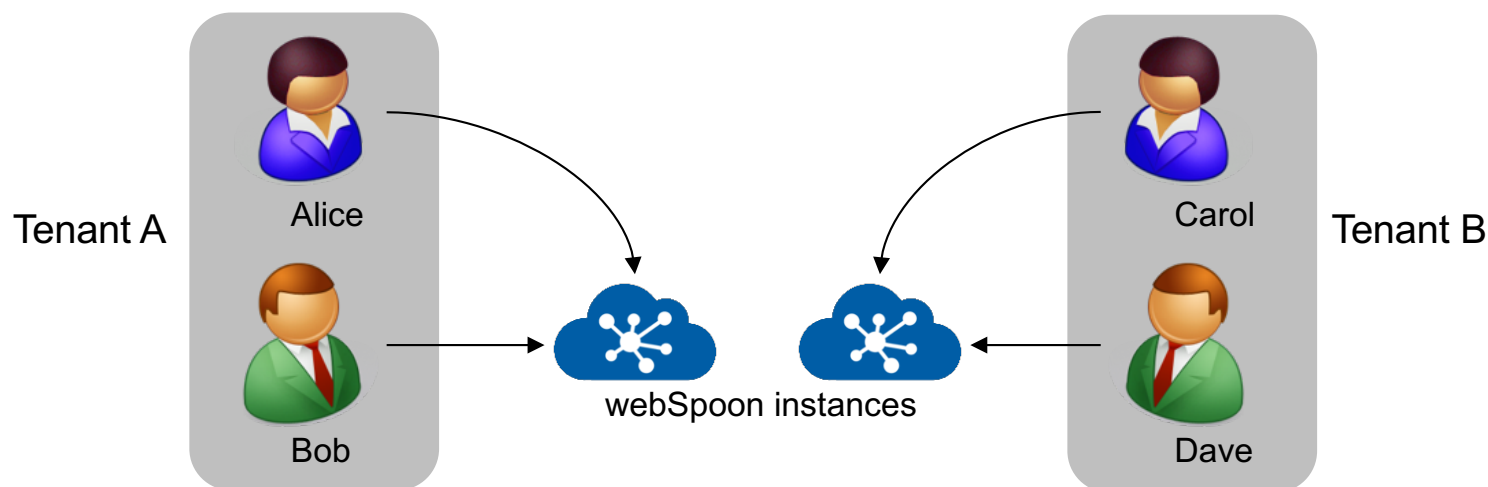
- webSpoon is scalable



- webSpoon serves multiple users
 - ✓ User authentication
 - ✓ User configuration
 - ✗ Incomplete privacy among users
 - Alice can see Bob's configuration files
 - Alice can see Bob's Kettle files (only when they are locally stored)
- webSpoon serves multiple tenants
 - I'd assign dedicated instances for each tenant for the privacy concern, though some argues that this arch is multi-instance, not multi-tenancy [1].



[1] Krebs, Rouven (2012). "Architectural Concerns in Multi-tenant SaaS Applications". Proc. 2nd Int. Conf. on Cloud Computing and Services Science (CLOSER 2012).



Compatible with Python/R steps

- Most of steps/job entries have been confirmed to be compatible with webSpoon, including

- Python (CPython Script Executor)
- R (Execute R Script)
- R (R script executor, EE only)

The screenshot shows the webSpoon interface. The top part displays a workflow diagram with two 'CSV file input' steps feeding into an 'R script executor' step. The bottom part shows the 'Execution Results' table with columns for row number, sepal length, sepal width, petal length, petal width, and class. The table contains four rows of data.

#	sepalength	sepalwidth	petalength	petalwidth	class
1	5.1	3.5	1.4	0.2	level1
2	4.9	3.0	1.4	0.2	level1
3	4.7	3.2	1.3	0.2	level1
4	4.6	3.1	1.5	0.2	level1

- The rest of steps/job entries is just left un-tested.

What's still missing?

- Security
 - End-users inherit the privileges of the user who runs the Tomcat.
 - If root runs the Tomcat, all end-users have the root permission.
 - Incomplete privacy among end-users:
 - Alice can see Bob's configuration files.
 - Alice can see Bob's Kettle files (when they are locally stored).
- Integration with Pentaho Server
 - Not realized yet due to un-resolved conflicts.
- Some EE features
 - DET (Data Exploration Tool)



Use cases

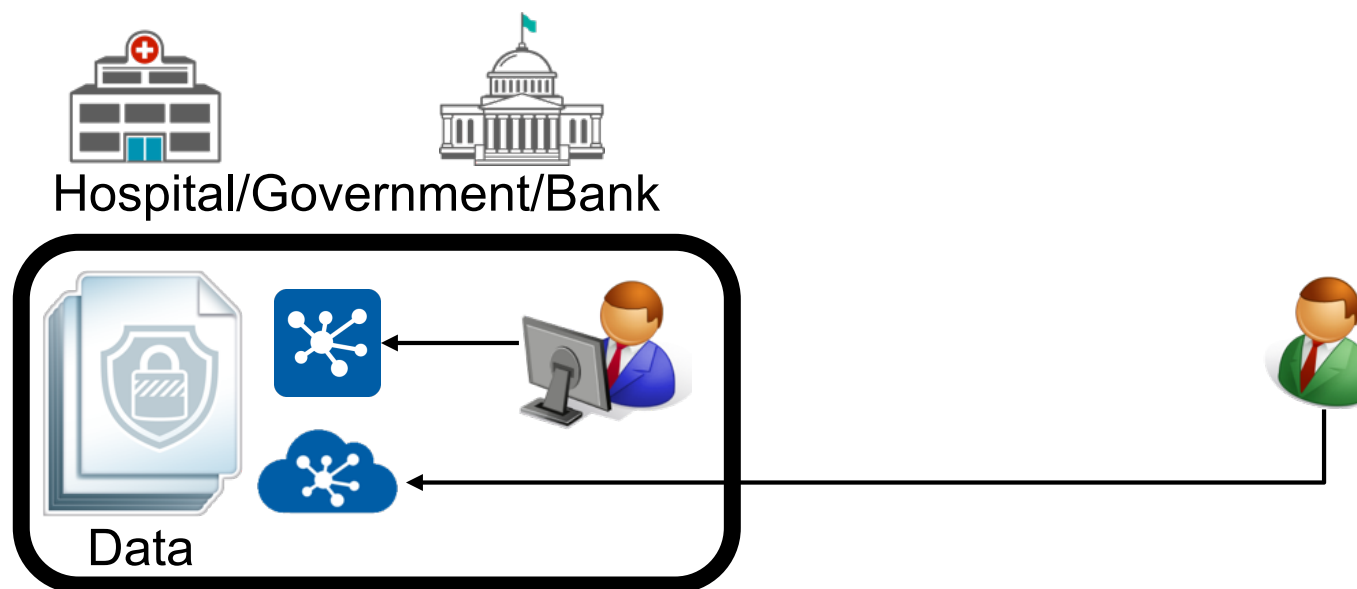
When data cannot leave facility/country due to some regulations,

Spoon

- Data engineers should physically be near data.
- They might be tempted to download data to work in their office.

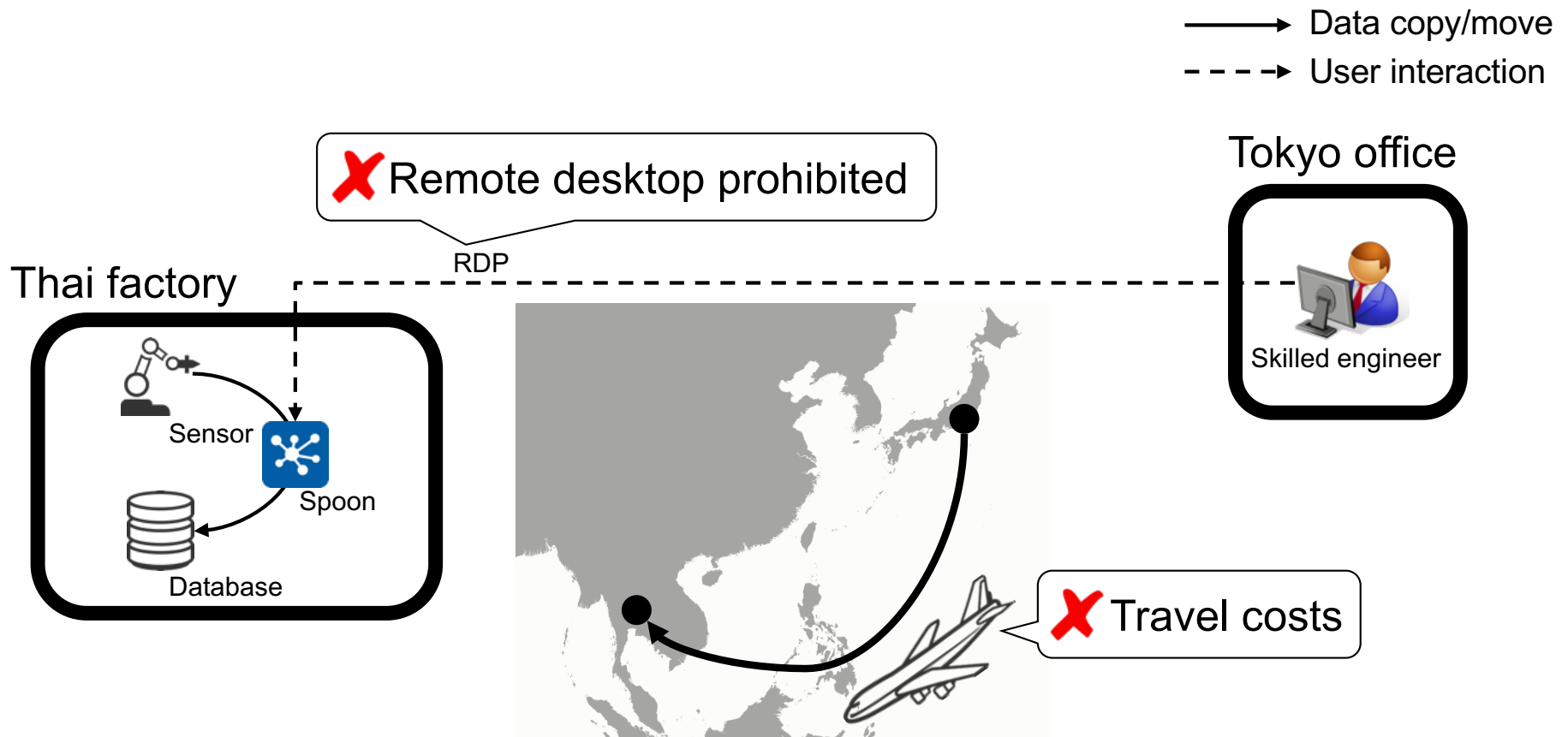
webSpoon

- They can work from office, home, or wherever comfortable.



Data integration of sensor data in remote sites

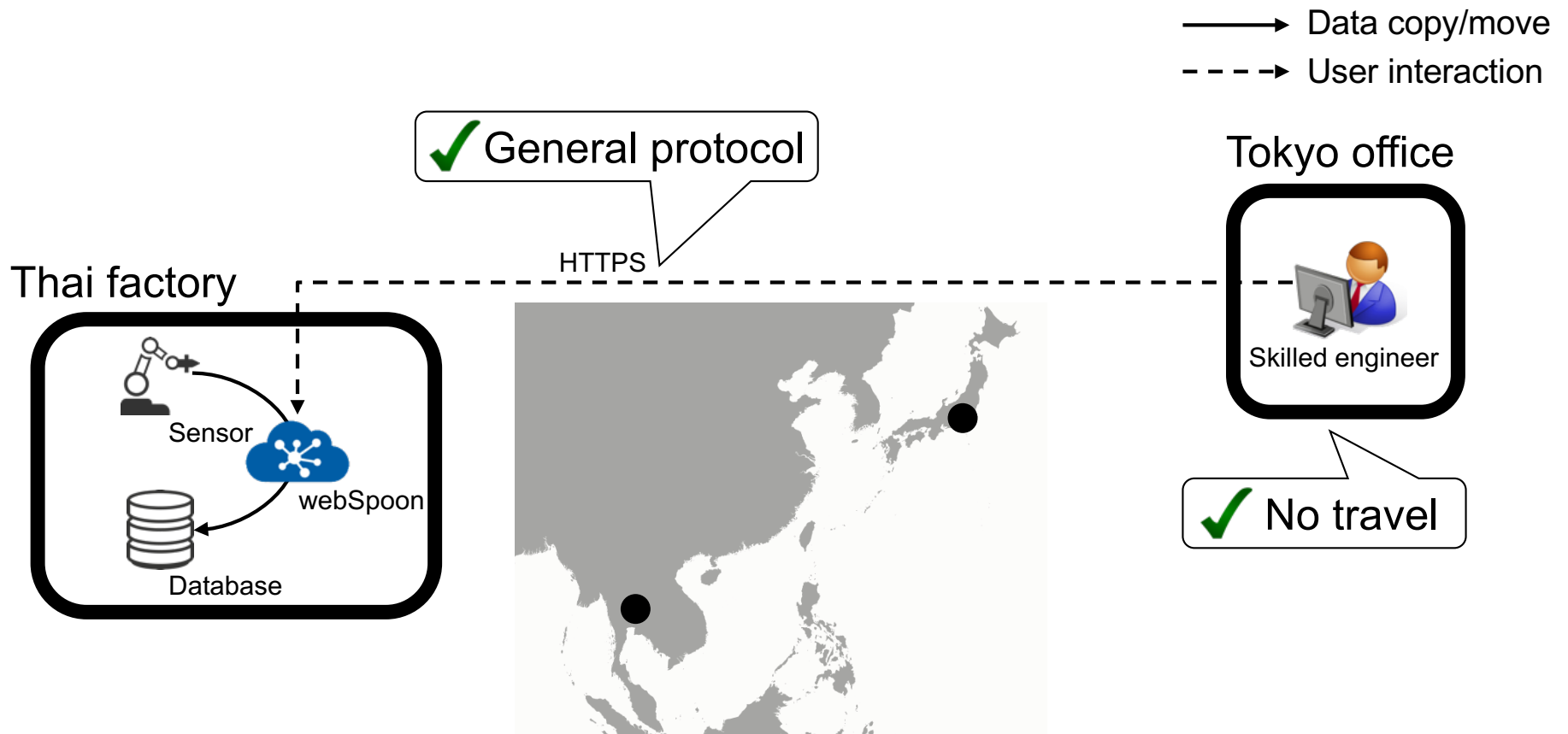
1. Kettle files need updating frequently for many reasons:
 - New machine, new sensor, new analytics, etc.
2. But, remote desktop (RDP) is prohibited and travel costs.



(Cropped) [Asia - Single Color](#) by [FreeVectorMaps.com](#)

Data integration of sensor data in remote sites

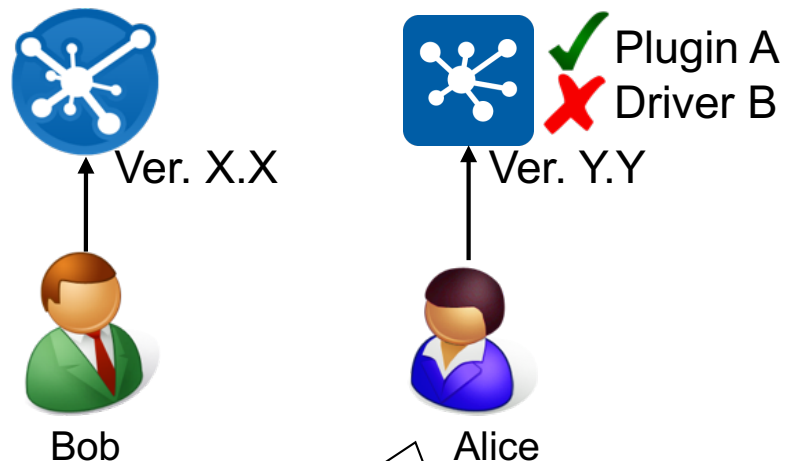
1. Kettle files need updating frequently for many reasons:
 - New machine, new sensor, new analytics, etc.
2. But, remote desktop (RDP) is prohibited and travel costs.



(Cropped) [Asia - Single Color](#) by [FreeVectorMaps.com](#)

Spoon

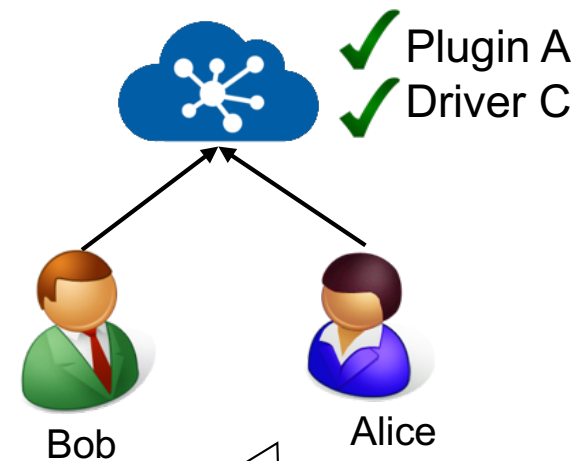
- Different version, plugin, etc. slows down collaboration.
- Could possibly be
 - Outdated.
 - Malicious plugins & drivers.



Your Kettle file does not run in my environment!

webSpoon

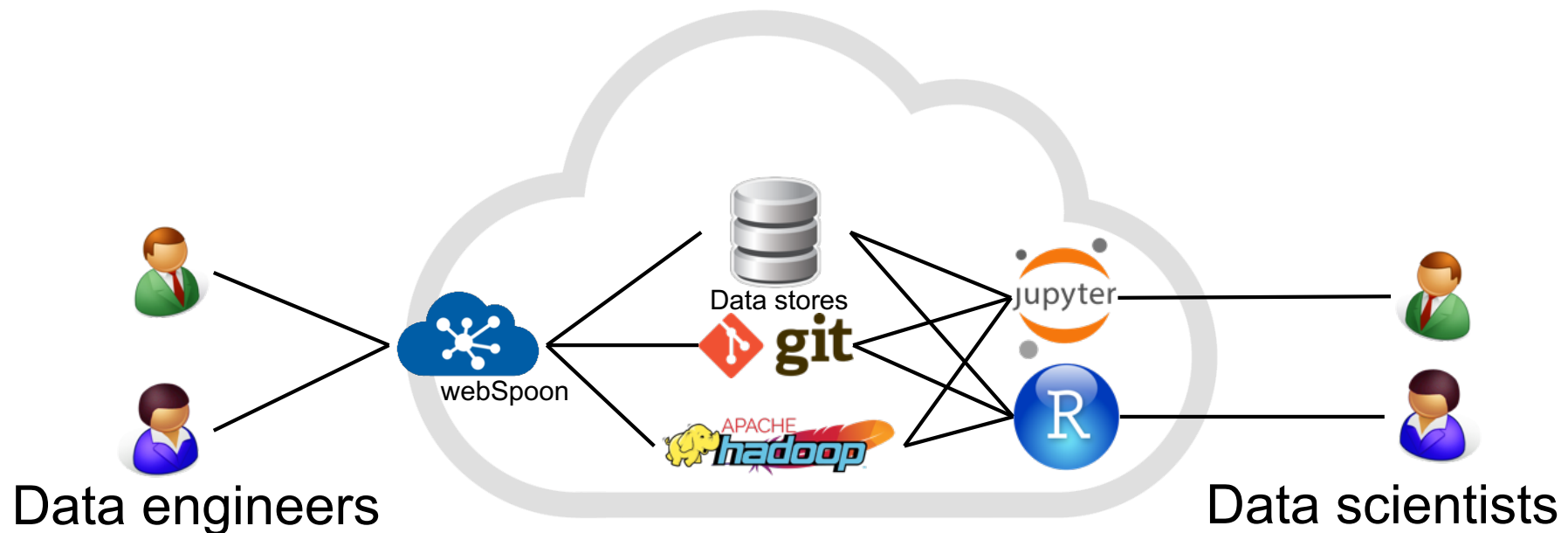
- All Kettle files run in coworker's screen.
- No installation/upgrade/update required (by end-users).
- Only desired plugins & drivers.



Your Kettle file runs in my environment!

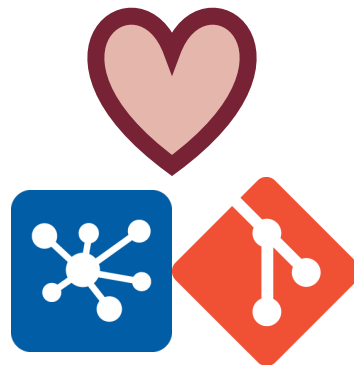
webSpoon streamlines the ML Workflow even more **HITACHI** Inspire the Next

- Data engineers/scientists share
 - Tools (Pentaho/Python/R)
 - Data stores
 - Git repository
 - Computing resources (e.g., Hadoop, Spark)
- As a result, collaboration between them becomes even more seamless
 - Less dependent on IT staffs to setup tools, data stores, etc.
 - No data copy/movement, no data dispersion



- Source and binary
 - <https://github.com/HiromuHota/pentaho-kettle>
- Docker image
 - <https://hub.docker.com/r/hiromuhota/webspoon>

One more thing...



SpoonGit (Git client integrated with Spoon)

The screenshot shows the SpoonGit application window titled "Spoon - Welcome!". The interface includes a toolbar with icons for repository management and a menu bar with options like "Git Repository", "Config", "Pull", "Push", "Branch", "Tag", "Refresh", and "testrepo / master".

Id	Message	Author	Date
WORKI...	// WORKINGTREE	*	2017-10-31T1...
31c42f...	Another commit 2	Hiromu Hota	2017-10-25T1...
7541c...	Make another commit	Hiromu Hota	2017-10-25T1...
b993c...	Conflicted commit	Hiromu Hota	2017-10-25T0...
50b74...	Another commit	Hiromu Hota	2017-10-25T0...
02081...	Commit	Hiromu Hota	2017-10-25T0...
d0030...	Change on hoge	Hiromu Hota	2017-10-17T1...
516cd...	More change on master	Hiromu Hota	2017-10-17T1...
72a56...	Change on master	Hiromu Hota	2017-10-17T1...
0304cf...	Merge commit '76efd733df353aac1...	IEUser	2017-10-16T1...

```
diff --git a/Untitled.ktr b/Untitled.ktr
index 703acc3..b91eb10 100644
--- a/Untitled.ktr
+++ b/Untitled.ktr
@@ -438,8 +438,8 @@
  <order>
  </order>
  <step>
-  <name>Select values</name>
-  <type>SelectValues</type>
+  <name>CSV file input</name>
+  <type>CsvInput</type>
  <description />
  <distributed>Y</distributed>
  <custom_distribution />
@@ -448,8 +448,31 @@
  <method>none</method>
  <schema_name />
  </partitioning>
+  <filename />
+  <filename_field />
+  <rownum_field />
+  <include_filename>N</include_filename>
+  <separator>,</separator>
+  <enclosure>"</enclosure>
+  <header>Y</header>
```

Changed files: Untitled.ktr

Commit Message:
Another commit 2

Author: Hiromu Hota <hiromu.hota@hal.hitachi.com>

Commit

- Source and binary
 - <https://github.com/HiromuHota/pdi-git-plugin>
- Binary
 - Pentaho Marketplace (in preparation)

- Pentaho is a trademark registered by Hitachi Vantara.
- Apache Hadoop and its logo are either registered trademarks or trademarks of the Apache Software Foundation (ASF).
- Apache Spark, Spark and the Spark logo are trademarks of ASF.
- The Git Logo by [Jason Long](#) is licensed under the [Creative Commons Attribution 3.0 Unported License](#).
- The R logo is © 2016 The R Foundation.
- RStudio and the RStudio logo are all registered trademarks of RStudio.
- The Python logo is a trademark of the Python Software Foundation.
- Jupyter and the Jupyter logs are trademarks of the NumFOCUS foundation.
- Docker and the Docker logo are trademarks or registered trademarks of Docker, Inc. in the United States and/or other countries.
- The Jenkins logo is licensed under the [Creative Commons Attribution-ShareAlike 3.0 Unported License](#).
- GitHub is a trademark registered in the United States by GitHub, Inc.
- Other company and product names mentioned in this document may be the trademarks of their respective owners.

HITACHI
Inspire the Next 



Appendix

webSpoon = Spoon - SWT + RWT

- Spoon relies on SWT for UI widgets (e.g., button, dialog, canvas).
- RWT is a web alternative to SWT and “largely” implements SWT APIs, meaning Spoon can become a web app with most codes intact.
- There are
 - Unimplemented SWT APIs (e.g., a part of GC, some Mouse events)
 - RWT-specific additional APIs (e.g., Multi-user, File Up/Download).

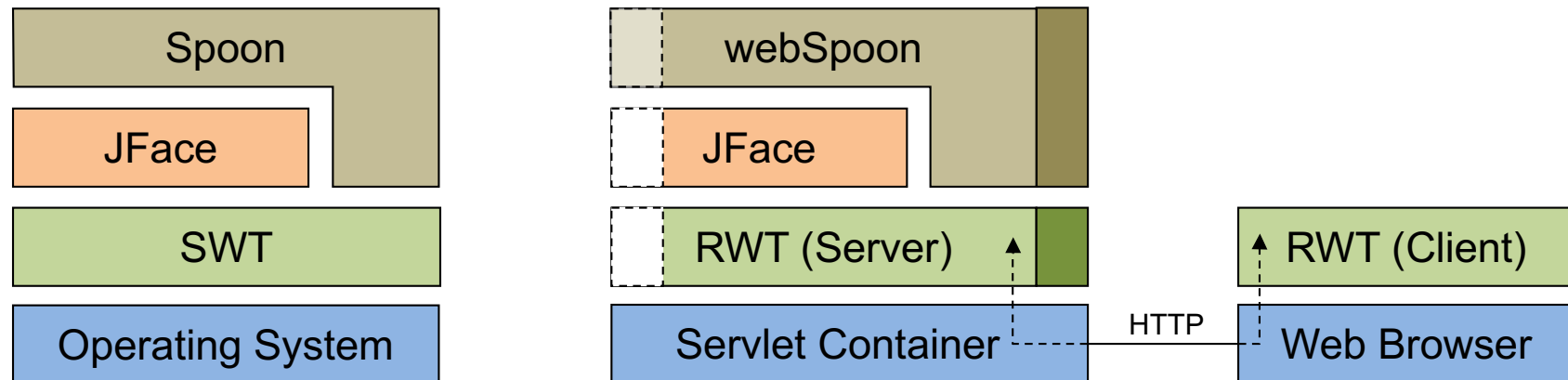
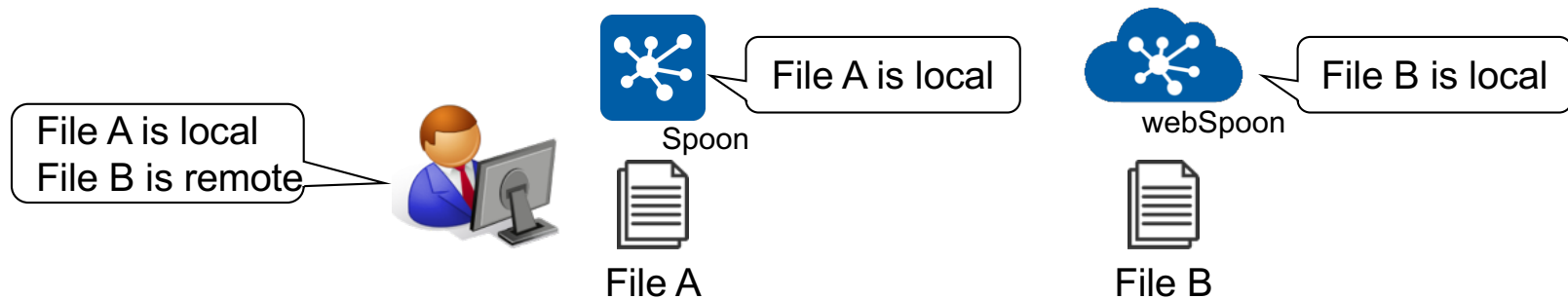


Image adapted from https://angelozerr.wordpress.com/2011/05/24/rap_step5/

How is webSpoon different from Spoon?

1. Local files

- Spoon: local files of the laptop/desktop
- webSpoon: local files of the (remote) server



2. Clipboard

- Spoon and webSpoon do not share the clipboard.
- In other words, no copy & paste between Spoon and webSpoon.

