Polly: A Language-Based Approach for Change Detection in Web Service Data

Elyas Ben Hadj Yahia CProDirect Jean-Rémy Falleri University of Bordeaux Laurent Réveillère University of Bordeaux

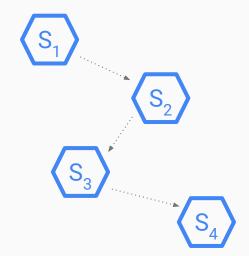
16/11/17 - ICSOC - Málaga





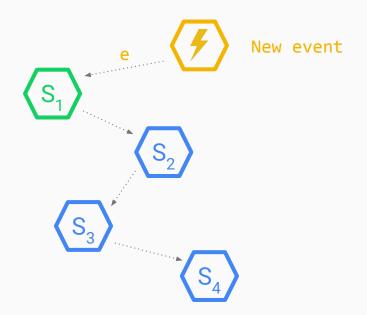


Service composition



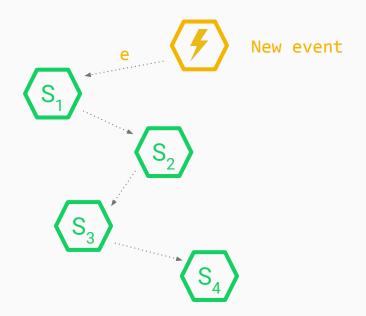
Orchestrating applications and services together to automate a **business process**

Service composition



Orchestrating applications and services together to automate a **business process**

Service composition



Orchestrating applications and services together to automate a **business process**





Customizable events



• "New questions on Stackoverflow"





- "New questions on Stackoverflow"
 - "... about Angular..."





- "New questions on Stackoverflow"
 - "... about Angular..."
 - "... with over **50+ bounty points**"

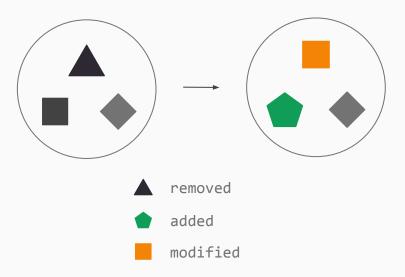




- "New questions on Stackoverflow"
 - "... about Angular..."
 - "... with over **50+ bounty points**"
- "A stock market share price decreased by 5+%"



- "New questions on Stackoverflow"
 - "... about Angular..."
 - "... with over **50+ bounty points**"
- "A stock market share price decreased by 5+%"
- "A change in the top 3 best-selling products"







Customizable events

Fast integration of new service providers

- "New questions on Stackoverflow"
 - "... about Angular..."
 - "... with over **50+ bounty points**"
- "A stock market share price decreased by 5+%"
- "A change in the top 3 best-selling products"

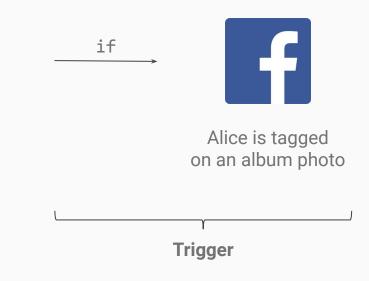


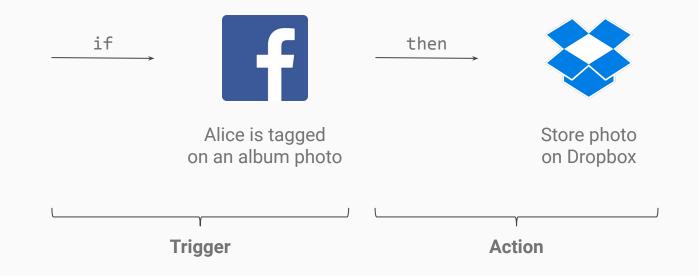
Customizable events

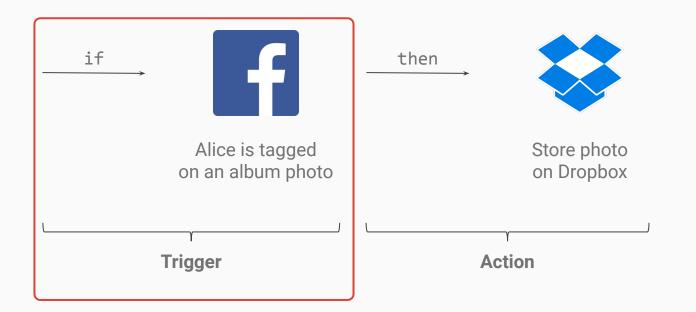
- "New questions on Stackoverflow"
 - "... about Angular..."
 - "... with over **50+ bounty points**"
- "A stock market share price decreased by 5+%"
- "A change in the top 3 best-selling products"

Fast integration of new service providers

- Rapidly support any RESTful web API
- Enable custom change detection strategies
- High-level abstraction of engineering concerns:
 - Data collection
 - Security, authentication
 - Maintenance, evolution







How to implement such triggers?

- Service providers only expose raw data (i.e. the state of a resource at a given time)
 - Example: Facebook only provides the following two API endpoints:

• Service providers only expose raw data (i.e. the state of a resource at a given time)

• Example: Facebook only provides the following two API endpoints:

GET /:albumId/photos

Retrieves the photos of an album on Facebook.



• Service providers only expose raw data (i.e. the state of a resource at a given time)

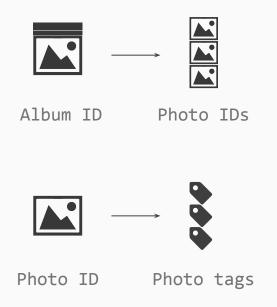
• Example: Facebook only provides the following two API endpoints:

GET /:albumId/photos

Retrieves the photos of an album on Facebook.

GET /:photoId/tags

Retrieves the users tagged in the photo.



GET /:albumId/photos





Album ID

GET /:albumId/photos







Album ID

Photo IDs

GET /:albumId/photos





Album ID

Photo IDs

GET /:albumId/photos



GET /:photoId/tags









Album ID

Photo IDs

Photo tags

```
"created_time": "2016-05-20T12:26:57",
"updated_time": "2016-05-20T12:28:57",
"id": "1106290499393017",
"data": [
    "id": "10203528656797589",
    "name": "Bob",
    "y": 74.865350089767
```

```
"created time": "2016-05-20T12:26:57",
"updated time": "2016-05-20T12:28:57",
"id": "1106290499393017",
"data": [
    "id": "10203528656797589",
    "name": "Bob",
    "x": 73.684210526316,
    "y": 74.865350089767
```

Initial state (t_o)



Updated state (t_1)





Initial state (t_0)

• Resulting JSON diff

```
"op": "replace",
  "path": "/0/data/0/x",
  "value": 76.684210526316
},
  "op": "replace",
  "path": "/0/updated_time",
  "value": "2016-05-20T12:29:57"
},
  "op": "add",
  "path": "/1",
  "value": {
    "created_time": "2016-05-20T12:35:57",
   "id": "2206280499393006",
    "data": [
        "id": "0406528656797578",
        "name": "Alice",
        "x": 63.684210526316,
        "y": 62.865350089767
```

• Resulting JSON diff

```
"op": "replace",
  "path": "/0/data/0/x",
  "value": 76.684210526316
了,
  "op": "replace",
  "path": "/0/updated_time",
  "value": "2016-05-20T12:29:57"
},
  "op": "add",
  "path": "/1",
  "value": {
    "created_time": "2016-05-20T12:35:57",
    "id": "2206280499393006",
    "data": [
        "id": "0406528656797578",
        "name": "Alice",
        "x": 63.684210526316,
        "y": 62.865350089767
```

Irrelevant changes (wrt. scenario)

• Resulting JSON diff



Irrelevant changes (wrt. scenario)

Relevant changes (wrt. scenario)

• Resulting JSON diff

Developers need to:

- Post-process the diff output
- Identify relevant changes
- Extract and transform them into a usable format

```
"op": "replace",
  "path": "/0/data/0/x",
 "value": 76.684210526316
ſ
 "op": "replace",
 "path": "/0/updated time",
 "value": "2016-05-20T12:29:57"
 "op": "add",
 "path": "/1",
 "value": {
   "created time": "2016-05-20T12:35:57",
   "id": "2206280499393006",
   "data": [
       "id": "0406528656797578",
       "name": "Alice",
        "y": 62.865350089767
```

Irrelevant changes (wrt. scenario)

Relevant changes (wrt. scenario)



• Challenges in polling for changes



State construction



Change detection

Chaining multiple endpoints and navigating through responses to construct a state

Identifying relevant changes, and ignoring irrelevant ones

Enter Polly



• Polly, a declarative DSL for custom change detection in web service data



• Polly, a declarative DSL for custom change detection in web service data



Language constructs for constructing a state from one or multiple API endpoints

Approach

• Polly, a declarative DSL for custom change detection in web service data



Language constructs for constructing a state from one or multiple API endpoints



Describing custom change detection strategies in JSON data from REST APIs

Approach

• **Polly**, a declarative DSL for custom change detection in web service data



Language constructs for constructing a state from one or multiple API endpoints



Describing custom change detection strategies in JSON data from REST APIs



Compiles to efficient Node.js code

Approach

• **Polly**, a declarative DSL for custom change detection in web service data



Language constructs for constructing a state from one or multiple API endpoints



Compiles to efficient Node.js code



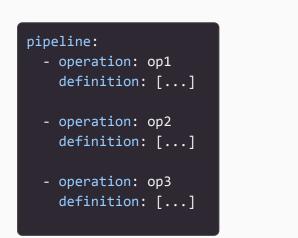
Describing custom change detection strategies in JSON data from REST APIs

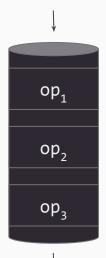


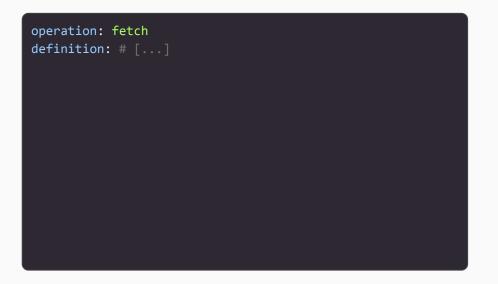
Validated on several real use-cases

Architecture

- Processing pipelines for defining custom change detectors
 - Expressed as a series of transformation operations on successive sets of data
 - \circ \quad Data and operations on it are independent from each other
 - Each operation produces a JSON document that is passed as input for the following operation







```
operation: fetch
definition:

request:
    url: https://graph.facebook.com/v2.9/:albumId/photos
    params:
        albumId: 465607303461343
    query:
        access_token: XXXX
    headers:
        Accept: application/json
```

- Request object construction
 - URL parameters
 - Query parameters
 - HTTP headers

```
operation: fetch
definition:
  request:
    url: https://graph.facebook.com/v2.9/:albumId/photos
    params:
        albumId: 465607303461343
    query:
        access_token: XXXX
    headers:
        Accept: application/json
    template: ~.data # ~ denotes the response body
```

- Request object construction
 - URL parameters
 - Query parameters
 - HTTP headers
- Custom templating
 - Transform response data
 - Include only relevant fields

```
operation: fetch
definition:
  request:
    url: https://graph.facebook.com/v2.9/:albumId/photos
    params:
        albumId: 465607303461343
    query:
        access_token: XXXX
    headers:
        Accept: application/json
        template: ~.data  # ~ denotes the response body
    pagination:
        next: ~.paging.next
```

- Request object construction
 - URL parameters
 - Query parameters
 - HTTP headers
- Custom templating
 - Transform response data
 - Include only relevant fields
- Automatic pagination handling
 - Recursively navigate to following pages

• Fetching multiple resources in parallel

```
operation: fetch
definition:
  request:
    url: https://graph.facebook.com/v2.9/:photoId/tags
    query:
      access token: XXXX
    headers:
      Accept: application/json
    pagination:
      next: ~.paging.next
    template:
      photoId: ^.id
      tags: ~.data
```

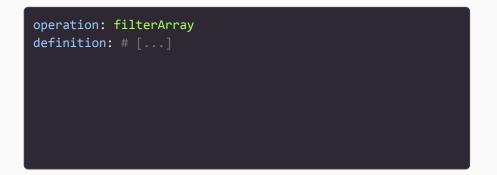
• Fetching multiple resources in parallel

operation: fetch definition:			
repeat:			
forEach: _	<pre># _ denotes the input object</pre>		
placeholders:			
photoId: ^.id	<pre># ^ denotes the iteration cursor</pre>		
request:			
url: https://graph	.facebook.com/v2.9/:photoId/tags		
query:	query:		
access_token: XX	XX		
headers:			
Accept: applicat	ion/json		
pagination:			
<pre>next: ~.paging.n</pre>	ext		
template:			
photoId: ^.id			
tags: ~.data			

- Multiple requests in parallel
 - Build a URL for each iteration cursor
 - Fetch resources asynchronously

Language operators - filterArray

• Detecting custom changes in array structures



Language operators - filterArray

• Detecting custom changes in array structures

	peration: filterArray	
ſ	identifiers:	
	- ^.photoId	

- Item identification
 - Specify how to uniquely identify an item within a collection

Language operators - filterArray

• Detecting custom changes in array structures

operation: filterArray	
definition:	
identifiers:	
- ^.photoId	
find:	
- addedItems	
output: &.addedItems	# & denotes the output object

- Item identification
 - Specify how to uniquely identify an item within a collection
- Change detection strategy
 - Change types to watch for
 - Leverage predefined change types

Language operators - filterCustom

• Custom filtering strategy

```
operation: filterCustom
definition:
  function: !!js/function >
    function (oldState, newState) {
      const result = [];
      # Filter the input, keeping only photos
      # where Alice was newly tagged.
      return result;
    }
```

Language operators - filterCustom

• Custom filtering strategy

```
operation: filterCustom
definition:
  function: !!js/function >
    function (oldState, newState) {
      const result = [];
      # Filter the input, keeping only photos
      # where Alice was newly tagged.
      return result;
    }
```

- User-defined logic
 - Javascript hook implementation
 - Runs in secure isolated sandbox

• Six use cases provided by our industrial partner

• Six use cases provided by our industrial partner



Notify when the top 5 best-selling products change in ranking order

• Six use cases provided by our industrial partner



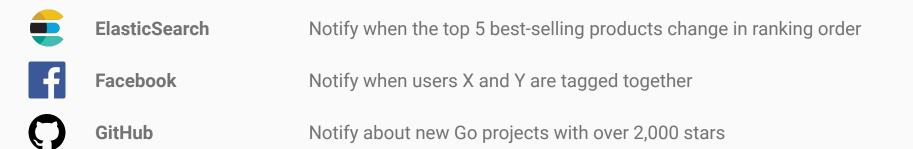
Facebook

ElasticSearch

Notify when the top 5 best-selling products change in ranking order

Notify when users X and Y are tagged together

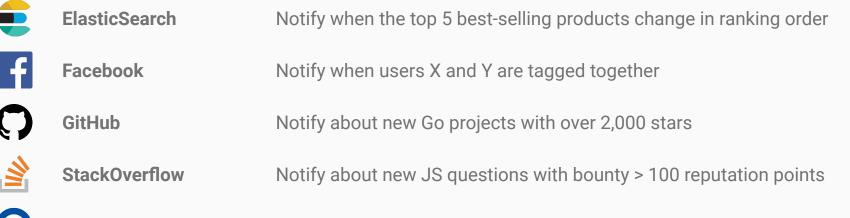
• Six use cases provided by our industrial partner



• Six use cases provided by our industrial partner

	ElasticSearch	Notify when the top 5 best-selling products change in ranking order
f	Facebook	Notify when users X and Y are tagged together
\mathbf{O}	GitHub	Notify about new Go projects with over 2,000 stars
	StackOverflow	Notify about new JS questions with bounty > 100 reputation points

• Six use cases provided by our industrial partner

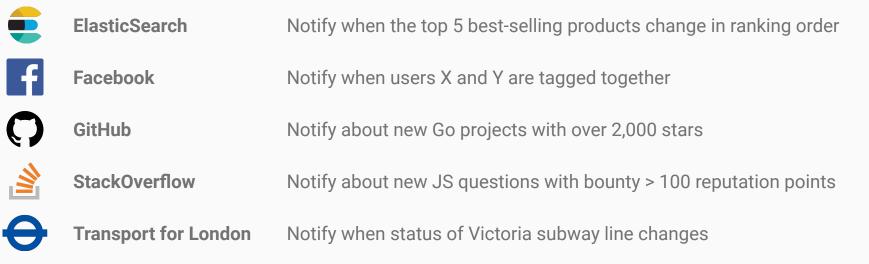




Transport for London

Notify when status of Victoria subway line changes

• Six use cases provided by our industrial partner



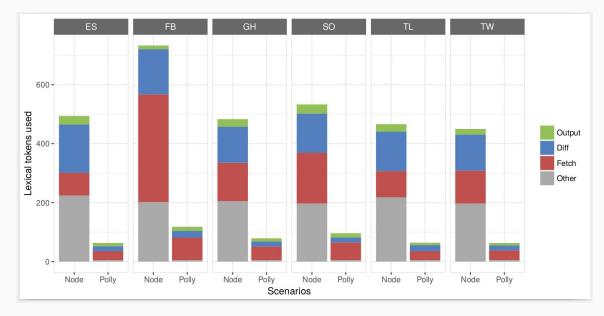


Twitter

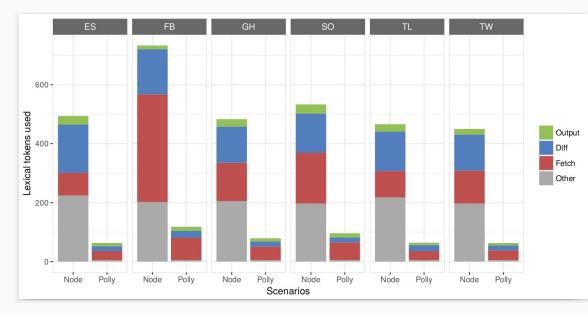
Notify whenever the official Bordeaux account has new followers

- Analysis of abstraction level
 - Verbosity measured by number of lexical tokens used to express a scenario
 - Comparison of handwritten Node.js implementation vs. Polly

- Analysis of abstraction level
 - Verbosity measured by number of lexical tokens used to express a scenario
 - Comparison of handwritten Node.js implementation vs. Polly

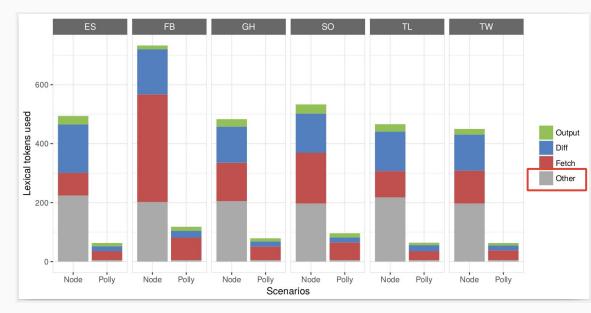


- Analysis of abstraction level
 - Verbosity measured by number of lexical tokens used to express a scenario
 - Comparison of handwritten Node.js implementation vs. Polly



 Programs written in Polly
 5.5 to 8 times smaller than Node.js

- Analysis of abstraction level
 - Verbosity measured by number of lexical tokens used to express a scenario
 - Comparison of handwritten Node.js implementation vs. Polly



 Programs written in Polly
 5.5 to 8 times smaller than Node.js

• Requires a lot less boilerplate code

- Performance analysis
 - Comparison against the top performing JSON differencing tool: JDR
 - Analyze differencing time and output size

- Performance analysis
 - Comparison against the top performing JSON differencing tool: JDR
 - Analyze differencing time and output size

• Experimental setup



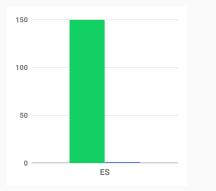
Collect real data from services (snapshot every 5 min for 48 hours = 576 snapshots)

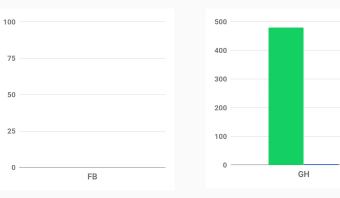


Mock server Serve collected data

Run scenarios and measure metrics

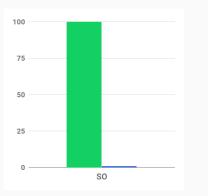
Evaluation - Output size

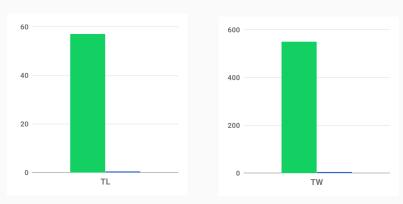




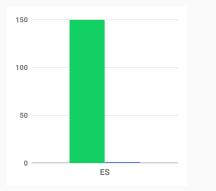


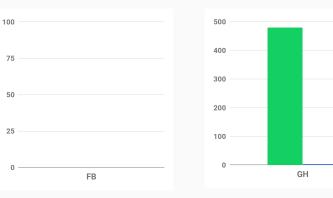
• Median values of output sizes (in bytes) per case





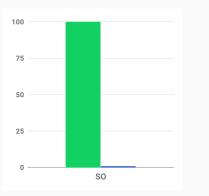
Evaluation - Output size

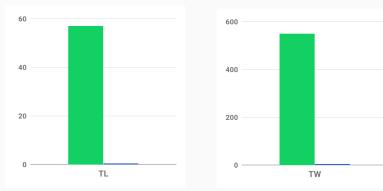






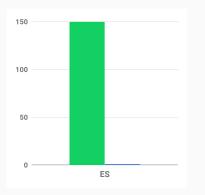
• Median values of output sizes (in bytes) per case





• **Polly** outperforms **JDR** in all cases

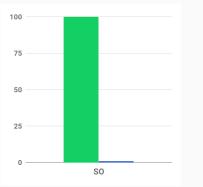
Evaluation - Output size

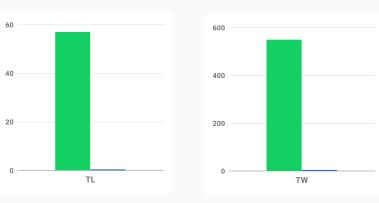






 Median values of output sizes (in bytes) per case

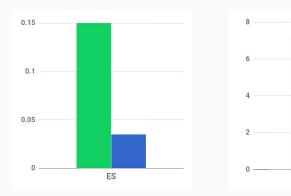


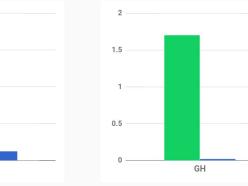


GH

- **Polly** outperforms **JDR** in all cases
- **FB** scenario: no changes during polling period

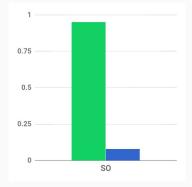
Evaluation - Differencing time

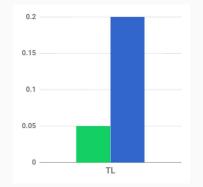




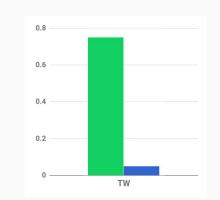


• Median values of diff time (in milliseconds) per case

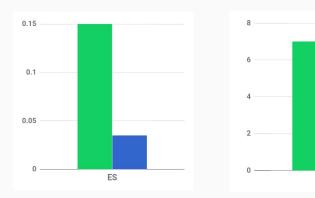


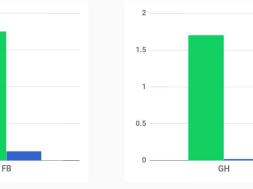


FB



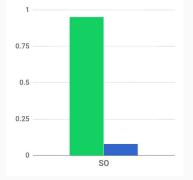
Evaluation - Differencing time

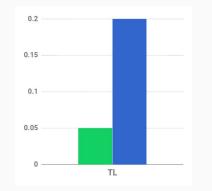


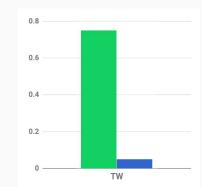




• Median values of diff time (in milliseconds) per case







 Polly outperforms JDR in almost all cases (0.15 ms slower for TL)

Statistical significance tests

- One-tailed paired Wilcoxon rank test
- Cohen's *d* level reported on Cohen's standard scale

Scenarios	Detection time (effect size)	Output size (effect size)
ES	Large	Medium
f FB	Large	NA
GH GH	Large	Large
🔌 SO	Large	Medium
	Large	Large
У ТW	Large	Large

Statistical significance tests

- One-tailed paired Wilcoxon rank test
- Cohen's *d* level reported on Cohen's standard scale

Scenarios	Detection time (effect size)	Output size (effect size)
ES ES	Large	Medium
f FB	Large	NA
GH	Large	Large
🔌 SO	Large	Medium
	Large	Large
y tw	Large	Large

 Polly results in a significantly improved outcome compared to JDR

Statistical significance tests

- One-tailed paired Wilcoxon rank test
- Cohen's *d* level reported on Cohen's standard scale

Scenarios	Detection time (effect size)	Output size (effect size)
ES	Large	Medium
f FB	Large	NA
GH GH	Large	Large
🔰 SO	Large	Medium
	Large	Large
🈏 тw	Large	Large

- Polly results in a significantly improved outcome compared to JDR
- **FB**: output size is equal to 0 for every polling step for both approaches.

Conclusion

• Polly, a declarative DSL for custom change detection in web service data



Language constructs for state construction from multiple API endpoints



Efficient implementation, outperforming existing solutions



Fine-grained custom change detection strategies



Validated on several real use-cases

Demo

• A demo site is available at the following address:

https://demo.pollyapp.ml

Polly: A Language-Based Approach for Change Detection in Web Service Data

Elyas Ben Hadj Yahia

elyas.bhy@cprodirect.fr

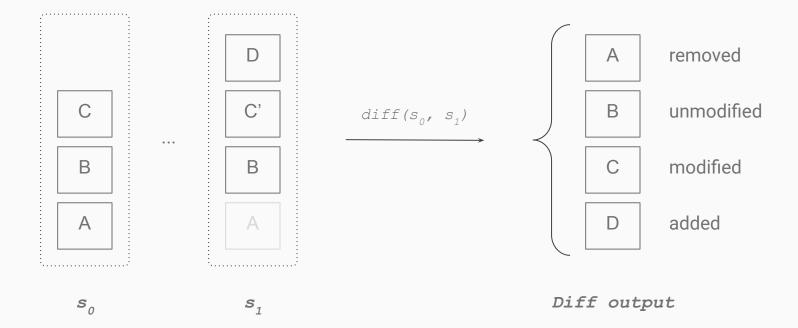
16/11/17 - ICSOC - Málaga







State differencing



• Webhooks



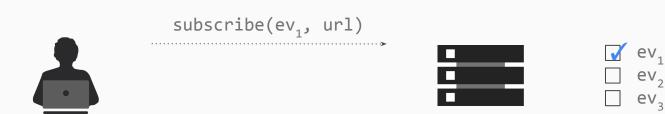


- Webhooks
 - Service providers exposes a set of predefined events





- Webhooks
 - Service providers exposes a set of predefined events
 - Client subscribes to the event(s) by providing a callback URL



- Webhooks
 - Service providers exposes a set of predefined events
 - Client subscribes to the event(s) by providing a callback URL
 - Service provider calls the provided URL when the specified event occurs

subscribe(ev, url) ev, ev, ev₃ notify(ev, payload)

- Webhooks
 - Service providers exposes a set of predefined events
 - Client subscribes to the event(s) by providing a callback URL
 - Service provider calls the provided URL when the specified event occurs

-	<pre>subscribe(ev₁, url)</pre>	
	·····>	Vev1
		ev ₂
	<pre>notify(ev,, payload)</pre>	

- + Realtime notification
- + No wasted network roundtrips
- + Easy to setup

• Webhooks

- Service providers exposes a set of predefined events
- Client subscribes to the event(s) by providing a callback URL
- Service provider calls the provided URL when the specified event occurs



- + Realtime notification
- + No wasted network roundtrips
- + Easy to setup

- Provided by very few services in practice
- Limited to predefined events

• Example: GitHub repository webhooks

Webhooks / Add webhook

We'll send a POST request to the URL below with details of any subscribed events. You can also specify which data format you'd like to receive (JSON, x-www-formurlencoded, etc). More information can be found in our developer documentation.

ŧ

Payload URL *

https://example.com/postreceive

Content type

application/json

Which events would you like to trigger this webhook?

Just the push event.

Send me everything.

- Let me select individual events.
 - Commit comment Commit or diff commented on.
 - Delete Branch or tag deleted.

Issues

Issue opened, edited, closed, reopened, assigned, unassigned, labeled, unlabeled, milestoned, or demilestoned. Create
 Branch or tag created.

 Issue comment Issue comment created, edited, or deleted.

Git push to a repository.

• Example: GitHub repository webhooks

Webhooks / Add webhook

We'll send a POST request to the URL below with details of any subscribed events. You can also specify which data format you'd like to receive (JSON, x-www-formurlencoded, etc). More information can be found in our developer documentation.

Payload URL *		
https://example.com/postrecei	ve	
Content type		
application/json	\$	

Which events would you like to trigger this webhook?

- Just the push event.
- Send me everything.
- Let me select individual events.
 - Commit comment Commit or diff commented on.
 - Delete Branch or tag deleted.
 - Issues

Issue opened, edited, closed, reopened, assigned, unassigned, labeled, unlabeled, milestoned, or demilestoned. Create
 Branch or tag created.

 Issue comment Issue comment created, edited, or deleted.

Git push to a repository.

• Example: GitHub repository webhooks

Webhooks / Add webhook

We'll send a POST request to the URL below with details of any subscribed events. You can also specify which data format you'd like to receive (JSON, x-www-formurlencoded, etc). More information can be found in our developer documentation.

ŧ

Payload URL *

https://example.com/postreceive

Content type

application/json

Which events would you like to trigger this webhook?

Just the push event.

Send me everything.

Let me select individual events.

Commit comment Commit or diff commented on.

Delete

Branch or tag deleted.

Issues

Issue opened, edited, closed, reopened, assigned, unassigned, labeled, unlabeled, milestoned, or demilestoned. Create
 Branch or tag created.

 Issue comment Issue comment created, edited, or deleted.

Push

Git push to a repository.

86



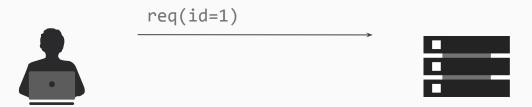
• Recurrent polling







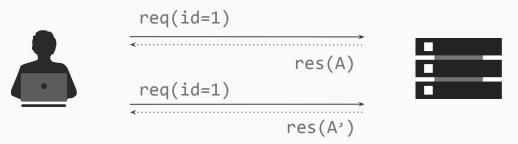
- Recurrent polling
 - Client polls a web service for a given resource



- Recurrent polling
 - Client polls a web service for a given resource
 - Service returns the current state of the requested resource

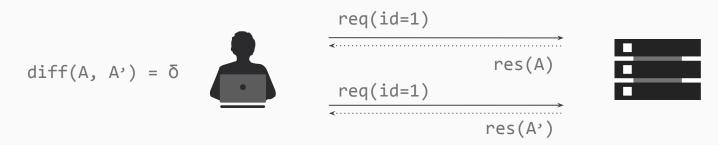


- Recurrent polling
 - Client polls a web service for a given resource
 - Service returns the current state of the requested resource
 - Client polls again later for the same resource



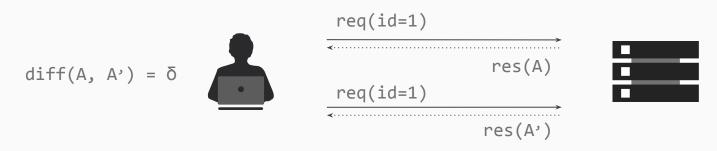
Recurrent polling

- Client polls a web service for a given resource
- Service returns the current state of the requested resource
- Client polls again later for the same resource
- Client calculates the difference between both states to detect any changes



Recurrent polling

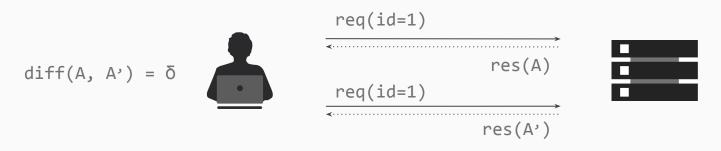
- Client polls a web service for a given resource
- Service returns the current state of the requested resource
- Client polls again later for the same resource
- Client calculates the difference between both states to detect any changes



+ Universal approach

Recurrent polling

- Client polls a web service for a given resource
- Service returns the current state of the requested resource
- Client polls again later for the same resource
- Client calculates the difference between both states to detect any changes



+ Universal approach

- Relatively complex to implement
- Tedious to support all event types