

Build A Serverless Data Pipeline

Lorna Mitchell, IBM



Stackoverflow Dashboard

The screenshot shows a web browser window displaying the Stackoverflow Dashboard. The browser's address bar shows the URL `https://sodashboard.mybluemix.net/#/all/-/-/-/`. The dashboard header includes navigation options like "SO Dashboard", "Owner", "Tags", and "Include" (with checkboxes for "Rejected" and "Answered"). A search bar is present with the placeholder text "Search for stuff!". The user's profile information shows a checkmark, the number "746", and the name "lornajane".

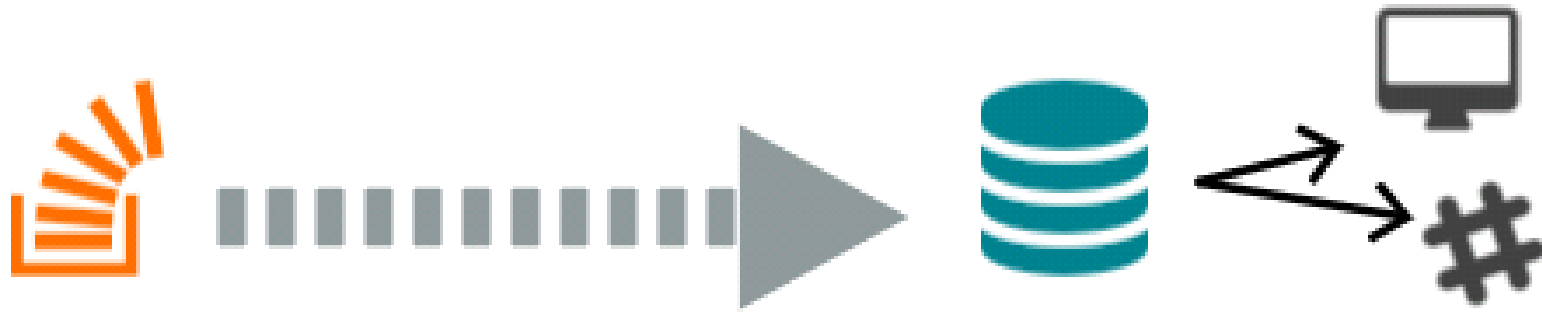
The main content area is titled "All Tickets" with a badge indicating 20 tickets. A "Show Notes" button is located to the right. The tickets are listed as follows:

Ticket Title	Author	Answers	Tags
How do I return Cloudant insert status?	H. Trujillo (62)	1 Answer	async-await, cloudant, node.js, promises
Cloudant search index response without one index field	S.Devnoza (1)	1 Answer	cloudant, indexing, lucene, search
how to list jars in the dsx spark environment and the jars loaded into the spark JVM?	Chris Snow (9216)	2 Answers	data-science-experience, ibm-bluemix, spark, jvm, jars
Loading .RData file into Data Science Experience	Venky (6)	0 Answers	data-science-experience, jupyter-notebook, r, notebook
posting text from DataFrame to IBM PersonalityInsights API	Brenzef (1)	1 Answer	api, data-science-experience, object-storage, personality-insights, python



Pipeline To Shift Data

Bringing data from StackOverflow into the dashboard my advocate team uses



What is Serverless?

Backend functions, deployed to the cloud, scaling on demand. Pay as you go.



The Serverless Revolution

FaaS: Functions as a Service

Developer focus:

- the outputs
- the inputs
- the logic in between

Charges are usually per GBsec



Why Go Serverless?

- Costs nothing when idle
- Small application, simple architecture
- Bursty usage since it runs from a cron
- No real-time requirement
- Easily within free tier

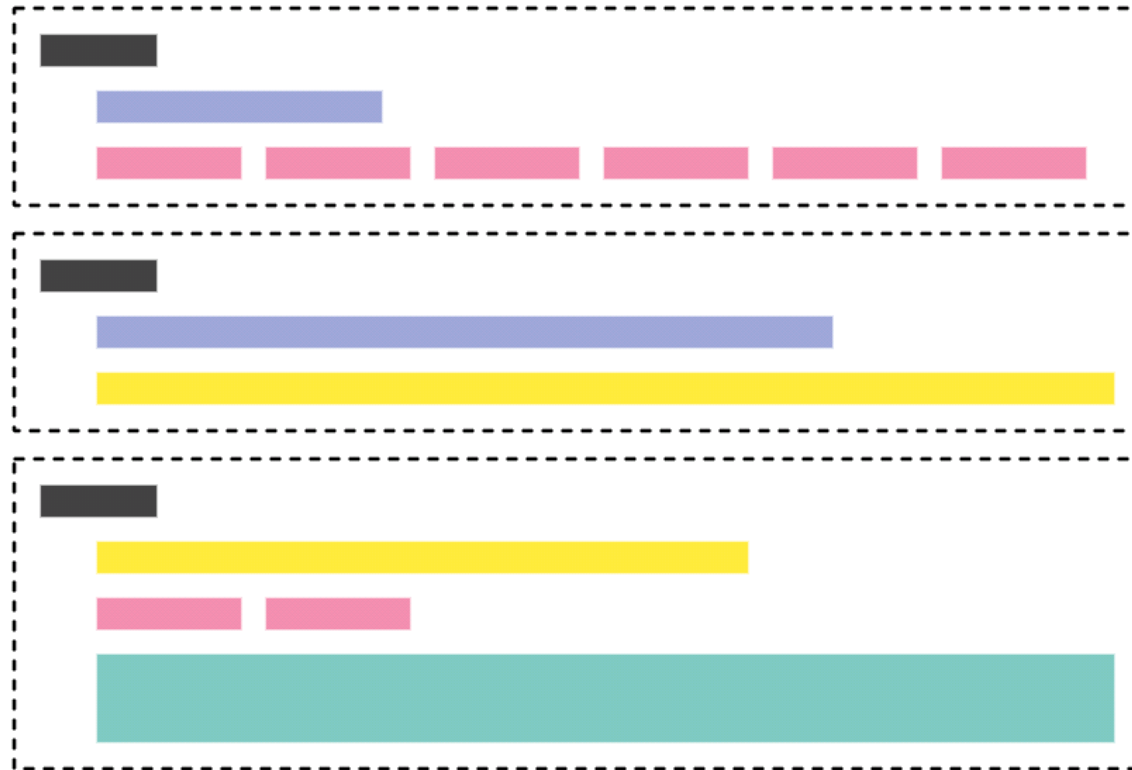


An Aside About Databases



Document Databases

Store collections of schemaless documents, in JSON



Apache CouchDB

Cluster of Unreliable Commodity Hardware

- Modern, robust, scalable document database
- HTTP API
- JSON data format
- Best replication on the planet (probably)



OfflineFirst Applications

This app is OfflineFirst:

- Client side JS
- Client side copy of DB using PouchDB
- Background sync to serverside CouchDB



Writing Serverless Functions



Serverless Platforms

- Amazon Lambda
- IBM Cloud Functions (aka OpenWhisk)
- Twilio Functions
- Azure Functions
- Google Cloud Functions
- ... and more every week

Hello World in JS

All the platforms are slightly different, this is for OpenWhisk

```
exports.main = function(args) {  
  return({"message": "Hello, World!"});  
};
```

Function must return an object or a Promise



OpenWhisk Vocabulary

- **trigger** an event, such as an incoming HTTP request
- **rule** map a trigger to an action
- **action** a function, optionally with parameters
- **package** collect actions and parameters together
- **sequence** more than one action in a row
- **cold start** time to run a fresh action



Working With Actions

Deploy code:

```
zip hello.zip index.js
```

```
bx wsk action update --kind nodejs:6 demo/hello1 hello.zip
```

Then run it:

```
bx wsk action invoke --blocking demo/hello1
```



Web-Enabled Actions

Deploy code:

```
zip hello.zip index.js
```

```
bx wsk action update --kind nodejs:6 --web true demo/hello1 hello.
```

Then curl it:

```
curl https://openwhisk.ng.bluemix.net/api/v1/web/.../hello1.json
```



Build the Data Pipeline



Designing for Serverless

- **Independent functions**
 - single purpose
 - testable
 - distributable



Start with Security

Need an API key or user creds for bx wsk tool

Web actions: we know how to secure HTTP connections, so do it!

- Auth standards e.g. JWT
- Security in transmission: use HTTPS



Logging Considerations

- Standard, configurable logging setup
- Use a `trace_id` to link requests between services
- Aggregate logs to a central place, ensure search functionality
- Collect metrics (invocations, execution time, error rates)
 - display metrics on a dashboard
 - have appropriate, configurable alerting



Pipeline Actions

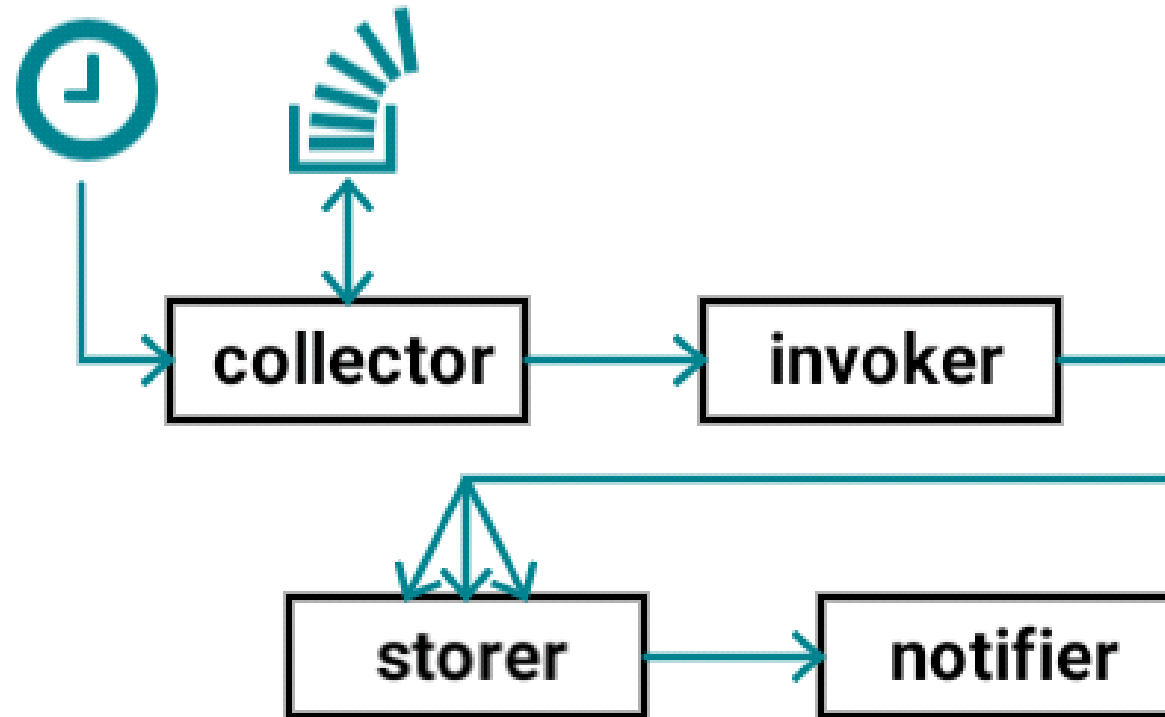
Sequence socron

- collector makes an API call, passes on data
- invoker fires many actions: one for each item

Sequence qhandler

- storer inserts or updates the record
- notifier sends a webhook to slack or a bot

Pipeline Actions



Serverless And Data



Resources

- Cloud Functions: <https://console.bluemix.net/openwhisk/>
- Code <https://github.com/ibm-watson-data-lab/soingest>
- My blog: <https://lornajane.net/>
- OpenWhisk: <https://openwhisk.org/>
- CouchDB: <https://couchdb.apache.org/>
- Offline First: <https://offlinefirst.org/>