

2016 dfrws europe

forensic analysis of cloud-native artifacts

VASSIL ROUSSEV

SHANE McCULLEY



THE UNIVERSITY of
NEW ORLEANS

context

≡ **driving problem**

‡ cloud *drive* acquisition & analysis

≡ **traditional solution**

‡ go to the client, look for the leftovers

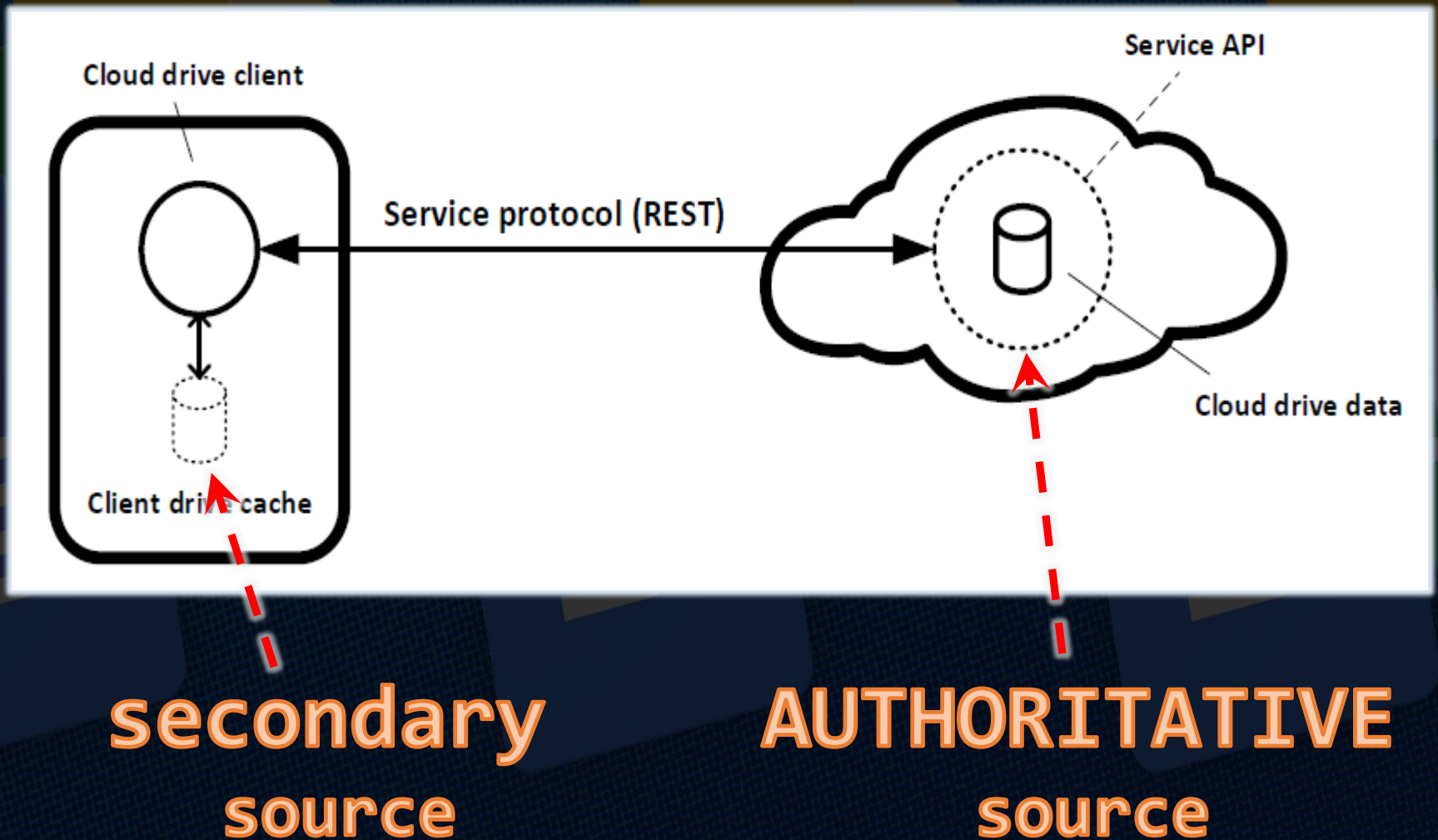
≡ **small detail** → it does **NOT** work:

‡ partial replication (data may not be on device)

‡ versions (only one on the client)

‡ cloud-native artifacts (e.g. Google Docs)

client acquisition is **un**sound





fixing it: kumodd/fs

≡ approach: use the API

≡ problems solved

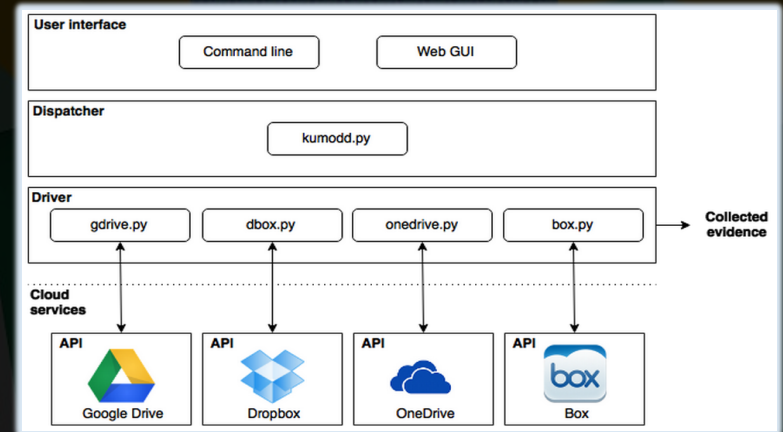
‡ partial replication ✓

‡ revision acquisition ✓

‡ cloud natives? ✓ ✗

≡ this work:

‡ provide a solution for cloud-native artifacts in Google Docs



cloud-native artifacts

≡ definition

- ‡ data objects which maintain the persistent state of web/SaaS applications,
 - ‡ and have no external representation on the client.
- ➔ these are *internal objects* for the app.

the start

NOVEMBER 5TH, 2014

How I REVERSE ENGINEERED GOOGLE DOCS

To Play Back Any Document's Keystrokes

JAMES SOMERS

Draftback

offered by jsomers.net

★★★★★ (33)

Productivity

20,368 *us*

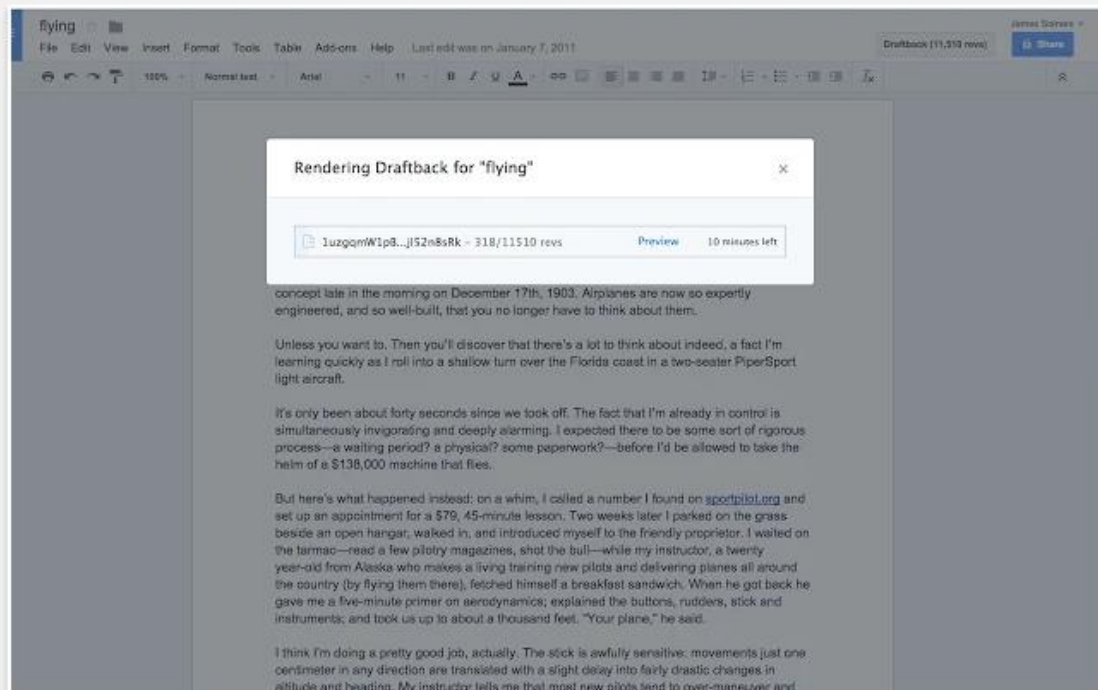
OVERVIEW

REVIEWS

2011-01-18

INTERVIEW

1000



Compatible with your device

The archaeology of great writing

Draftback lets you play back the revision history of any Google Doc you can edit. It's like going back in time to look over your own shoulder as you write. Notes:

- With Draftback, your data is kept entirely private. Draftback was purposely designed so that you could play back your own docs without having to share them with a third party. This is -your- data; Draftback just lets you see it in a new way.

- Draftback only needs access to docs.google.com to get the revision data for



 Website



 [Report Abuse](#)

Version: 0.0.8

Updated: March 2, 2015


Size: 471KB

Language: English (United States)

USERS OF THIS EXTENSION HAVE ALSO USED

revisions

Revision history
Today, 6:43 PM

 100% ▼ Total: 1 edit ^ ▼

Common mistakes from CSCI 4311 PA2:

~~Some test text~~

1a) Using available() to check for the end of a stream
Classes implementing available() return an “estimate” of the number of bytes that can be read without blocking. If this returns zero, it only means there is no data to read **now**, but there may be data to read in the future. Streams backed by sockets or some other source that buffers data often need more time for data to become available.

A more reliable indicator of the end of a stream is a return value of -1 from a read() method, or null from readLine(). This is the only guarantee that no more data will be available, aside from an IOException for some other reason.

1b) Using available() to check whether an input stream has been redirected
Unfortunately, there does not seem to be a portable way to check whether stdin has been redirected in Java. In C, the “isatty” function will let you check whether stdin or stdout is

Revision history

Today, 6:43 PM
■ Vassil Roussev

September 27, 8:32 PM
■ Vassil Roussev

March 18, 2013, 9:12 PM
■ Tom Sires

March 12, 2013, 1:10 AM
■ Tom Sires


March 12, 2013, 12:10 AM
■ Tom Sires

March 11, 2013, 10:59 PM
■ Tom Sires

March 11, 2013, 8:06 PM
■ Tom Sires

March 11, 2013, 7:24 PM
■ Tom Sires

March 11, 2013, 6:38 PM
■ Tom Sires

☒ Show changes 

Show more detailed revisions

more(!) revisions

CSCI-4311-PA2 Notes ☆ 📁

vassil@roussev.net ▾

File Edit View Insert Format Tools Table Add-ons Help Last edit was 24 min...

Draftback (2,971 revs)

Comments

Share

Rendering Draftback for "CSCI-4311-PA1 Notes"

📄 1EhfdnBKHw...QssfyN568 - 1717/1717 revs

[View](#)



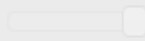
Common mistakes from CSCI 4311 PA2:

1a) Using available() to check for the end of a stream

Classes implementing available() return an "estimate" of the number of bytes that can be read without blocking. If this returns zero, it only means there is no data to read **now**, but there may

p(l)ayback

[Finish & Publish 387 revisions](#)



☐ play at actual speed?

Sun, 3/3/2013, 3:51:48 PM

[document graphs and statistics](#)

Common issues from CSCI 4311 PA1:

Using line-buffered I/O.
BufferedReader.readLine() A line is considered to be terminated by any one of a line feed ('`\n`'), a carriage return ('`\r`'), or a carriage return followed immediately by a linefeed.
For text data, this is generally not a problem, as long as platform-independent linebreaks are added to output, e.g. with `BufferedWriter.newLine()`.

Relevant docs:
[http://docs.oracle.com/javase/7/docs/api/java/io/BufferedReader.html#readLine\(\)](http://docs.oracle.com/javase/7/docs/api/java/io/BufferedReader.html#readLine())
[http://docs.oracle.com/javase/7/docs/api/java/io/BufferedWriter.html#newLine\(\)](http://docs.oracle.com/javase/7/docs/api/java/io/BufferedWriter.html#newLine())

Common issues from CSCI 4311 PA1:

Using line-buffered I/O.

BufferedReader.readLine() A line is considered to be terminated by any one of a line feed ('`\n`'), a carriage return ('`\r`'), or a carriage return followed immediately by a linefeed."

For text data, this is generally not a problem, as long as platform-independent linebreaks are added to output, e.g. with `BufferedWriter.newLine()`.

behind the scenes

≡ clearly, gDocs stores *everything* you did!

≡ why?

- ‡ why not?

 - ~ bandwidth & storage on the house!

- ‡ user analytics

 - ~ "if you are not paying ... you *are* the product"

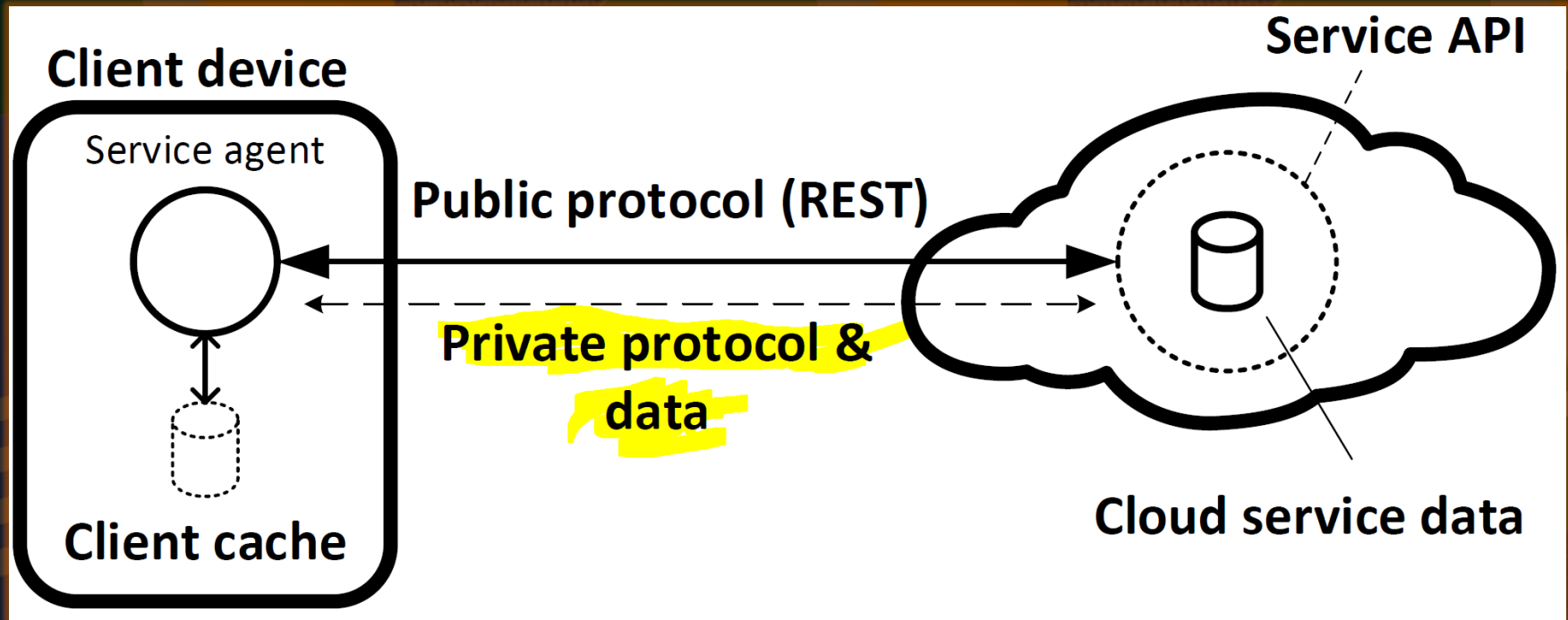
- ‡ user convenience?

 - ~ can you handle 10k revisions?

- ‡ programmer convenience?

 - ~ works with the real-time collaboration concept

web app architecture



meet the changelog

☐ load?id=1IdObEjEPRwAfmYoaSc6vVZVYgrM3oag_mU-Y-3Mj4FQ&start=4943&end=4960&token=AC4w5...

```
▼ changelog: [[{"ty": "mti", "mts": [{"ty": "ds", "si": 14776, "ei": 14776}, {"ty": "ds", "si": 14775, "ei": 14775}]],...],...]  
▶ 0: [{"ty": "mti", "mts": [{"ty": "ds", "si": 14776, "ei": 14776}, {"ty": "ds", "si": 14775, "ei": 14775}]],...]  
▶ 1: [{"ty": "mti", "mts": [{"ty": "ds", "si": 14774, "ei": 14774}, {"ty": "is", "s": ".", "ibi": 14774}]], 1443757842902,...]  
▶ 2: [{"ty": "is", "s": " ", "ibi": 14775}, 1443757843127, "18178839968700900856", 4945, "df36183a2f26250", 660,...]  
▶ 3: [{"ty": "is", "s": "Afte", "ibi": 14777}, 1443757843660, "18178839968700900856", 4946, "df36183a2f26250",...]  
▶ 4: [{"ty": "is", "s": "r a", "ibi": 14781}, 1443757843854, "18178839968700900856", 4947, "df36183a2f26250", 662,...]  
▶ 5: [{"ty": "is", "s": "n h", "ibi": 14784}, 1443757844440, "18178839968700900856", 4948, "df36183a2f26250", 663,...]  
▶ 6: [{"ty": "is", "s": "our ", "ibi": 14787}, 1443757844751, "18178839968700900856", 4949, "df36183a2f26250",...]  
▶ 7: [{"ty": "is", "s": "and ", "ibi": 14791}, 1443757845081, "18178839968700900856", 4950, "df36183a2f26250", 1
```


the chunkedSnapshot

```
"chunkedSnapshot": [
  [{"ty": "is", "s": "Test document", "ibi": 1},
  [{"ty": "as", "sm": {"hs_h1": {"sdef_ts": {"ts_fs": 18.0, "ts_fs_i": false}},
    ...
    "hs_h6": {"sdef_ts": {"ts_bd": false, "ts_bd_i": true, "ts_fg": "#666666"}},
    {"ty": "as", "sm": {"lgs_l": "en"}, "ei": 0, "st": "language", "fm": false, "si": 0},
    [{"ty": "as", "sm": {"ts_bd": false, "ts_bd_i": true, "ts_bgc": null, "ts_bgc_i": true, "ts_fg": "#000000", "ts_fg_i": true, "ts_fs": 11.0, "ts_fs_i": true, "ts_sc": false, "ts_sc_i": true, "ts_st": false, "ts_st_i": true, "ts_un": false, "ts_va": "nor", "ts_va_i": true}, "ei": 12, "st": "text", "fm": false, "si": 0}]}]
```

Drawing/Slides

```
{ "changelog": [
  [[3, "g27de7cf84_0_0", 108, [2.292, 0.0, 0.0, 0.2674, 63984.0, 37722.0], [44, 0, 45,
    1444063509783, "08413168629437028300", 2,
  [[15, "g27de7cf84_0_0", null, 0, "T"], 1444063511799, "08413168629437028300", 3,
  [[15, "g27de7cf84_0_0", null, 1, "e"], 1444063512119, "08413168629437028300", 4,
  [[15, "g27de7cf84_0_0", null, 2, "s"], 1444063512448, "08413168629437028300", 5,
  [[15, "g27de7cf84_0_0", null, 3, "t"], 1444063512448, "08413168629437028300", 6,
  [[3, "g27de7cf84_0_1", 99, [0.1432, 0.0, 0.0, 0.3263, 174285.0, 78309.0], [22, 381,
    1444063520352, "08413168629437028300", 12,
  , "chunkedSnapshot": [
    [[1, [365760, 274320], [302400, 427680]], [45, [], [0, "en"]], [13, 0, 0, null, "p"], [
    [12, "m", 0, 2, []], [12, "l", 0, 1, []], [12, "p", 0, 0, []]]
  ] }
```


embedded images

≡ upon upload

‡ a temporary Google CDN link is provided
(googleusercontent.com)

~ lasts ~1 hour

‡ a permanent CDN link is also provided

≡ CDN link is like a dead drop

‡ if you know the address, you can access it

~ no authentication

CDN meet changelog

- ≡ **what happens if we delete an image:**
 - a) CDN image is deleted and becomes unrecoverable, or
 - b) it's kept around, in case the change is rolled back?
- ≡ **well, b) of course!**
- ≡ **as long as *any* document revision references the image, the public CDN link will remain live!**
- ≡ **if the whole document is deleted, embeds are garbage collected**
 - ... after ~1 hour ...

reversions

Q: What happens when we revert to a prior version?

A:

- ‡ a snapshot of the desired revision is created,
- ‡ a “revert” entry w/ the snapshot is *appended* to the log.

observations & challenges [1]

≡ changelogs cannot be spoiled

‡ Google will only **add** things to the log; no way to permanently modify prior state

≡ **the golden CDN hour**

‡ could recover "SWAT-triggered" deletions

≡ **reverse engineering is still critical**

‡ but emphasis will shift to protocols

≡ **how representative is gDocs?**

‡ somewhat, at least

observations & challenges [2]

≡ collecting changelogs is easy ...

≡ storing & replaying them, much less so:

- ‡ what format should they be in?

- ‡ how do you render them (years from now)?

- ~ changelog is an internal data structure, it can change at *any* time

privacy

≡ how to edit-share a doc without sharing the history?

‡ you cannot → the history *is* the document

≡ workaround

‡ create a copy (zaps the history) and share that

≡ privacy audit

‡ extract all "deleted" embeds → make sure you still need them

≡ CDN link as a dead man's switch

‡ remember the librarian ...

contributions

≡ new problem formulation

‡ requires different methods & tools

≡ gDocs artifact & behavior analysis

‡ *Documents & Slides*

‡ protocol documentation

≡ PoC development: **kumodocs**

‡ github: [kumofx/kumodocs](https://github.com/kumofx/kumodocs) (coming soon)

thank you!

Questions?

vassil@roussev.net