
MAY 11-12, 2015 | SANTA CLARA, CA
CLOUD FOUNDRY
SUMMIT

Using Service Brokers to Manage Data Lifecycle

Josh Kruck | @krujos

jkruck@pivotal.io

github.com/krujos

Who am I?

- "Storage Industry Veteran"
  - 15 years doing stuff in storage, mostly around intersection of storage and apps.
- Symantec:
  - Information Management Tech Office
  - Architect @ Symantec's Appliance CoE
- Advisory Solutions Architect @ Pivotal
  - Customer success

What are the operational problems with data?

And why do they need managing?

Business Critical Data Lifecycle RTO 00:05 RPO 01:00 First 12 hours

Primary
Snapshots
Replica
Backup

Business Critical Data Lifecycle
RTO 00:05 RPO 01:00
First 12 hours

Primary

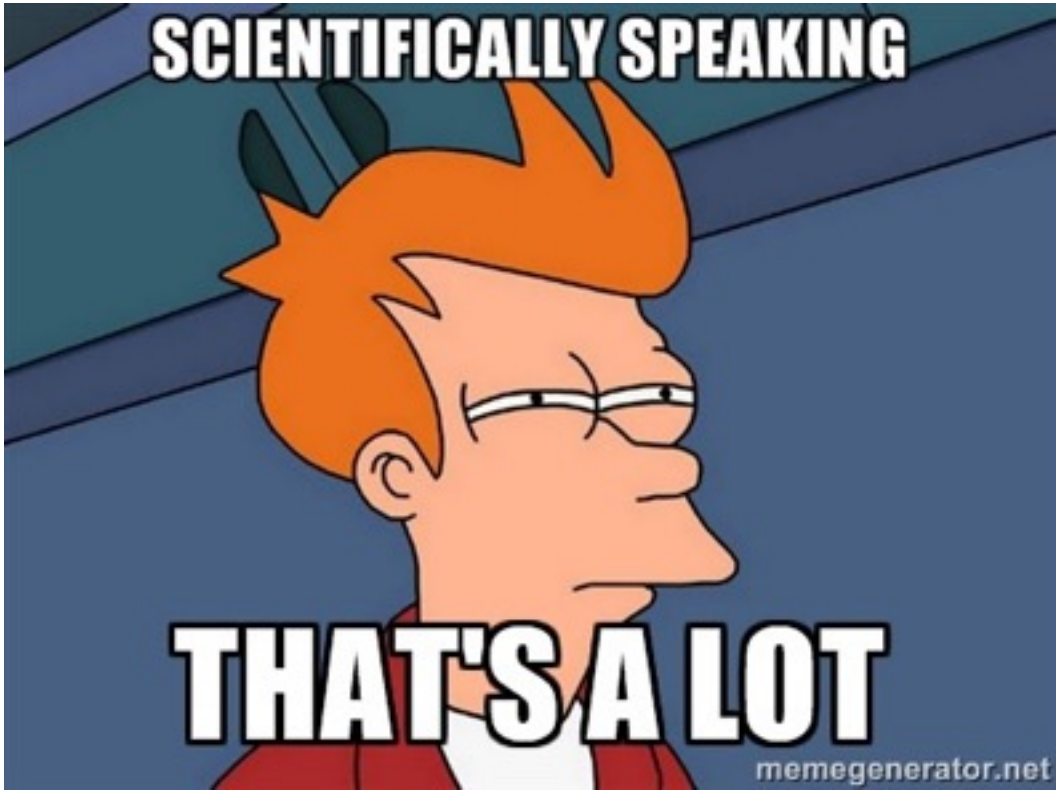
Snapshots

Replica

Backup

525,600 minutes

5476 copies


SCIENTIFICALLY SPEAKING
THAT'S A LOT
memegenerator.net

copies aren't really the problem!

(capex is easy, just buy more stuff)

The real problem is 5476 copies are…

managed by 3 systems

["storage", "backup", "rdbms"]

and 5 teams.

[
“storage”,
“backup”,
“offsite provider”,
“app owner”,
“dba”
]

opex is the problem

(you shouldn't buy more people)

fun facts

approximate read/write load on the all those copies?

0

5475 copies doing nothing for your business

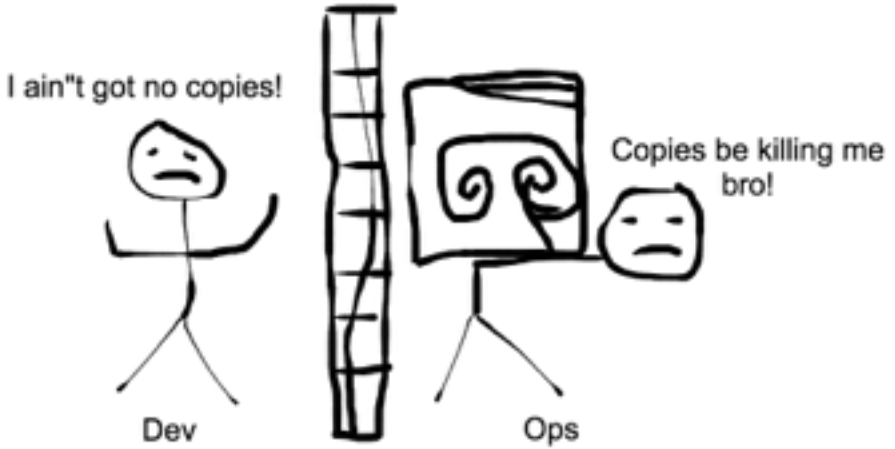
We're here to talk about CF, why all this talk about data and backups and stuff?

apps have data problems too!

As a developer I need the CI/CD pipeline to test against a current copy of production data so my changes can be deployed automatically with a high degree of confidence that my phone won't ring an hour later.

A play in 3 acts

“I don’t think we have any copies of that”



“I not allowed to have prod logs, much less the db”

okay, we can do it, this one time, through a ticket.

Solved!
But did we create another problem?

Once you find a copy, it needs a curator

- Sizing (don't use all of 10 TB of prod to test)
  - But your sample must represent the entirety of the dataset.
  - Efforts to size representative samples are usually futile (unknown unknowns).
  - Sizing means you restrict your tests to what you left in.
  - Sizing hides performance issues (missing index)
  - So maybe it's not worth it….

Once you find a copy, it needs a curator

Sanitize it! Can’t have SSN’s and CC in test

Once you find a copy, it needs a curator

Refresh it! Is stale data good data?

Once you find a copy, it needs a curator

- GOTO 10

A manual process that starts with a ticket is the wrong solution

Curation is expensive

It's manual

It's error prone

Who deletes copies

It's infrequent

It requires handoffs

Who is responsible

hard|complex

What happens if we combine the problems?

The sum of the mess is worth more than its parts

There’s 5475 secondary copies with no load, let’s use those for testing.

How?

first do no harm

most copies do nothing, but when the sky is falling you need them

Putting the E in Enterprise

- Buy a CDM Product
  - Actifio, Delphix, ViPR
  - Great if they support your workloads!
  - And you can consume the form factors they deliver

BYO

- Based on technology to allow layered writes
  - Layered FS (Docker, Docker, Docker)?
  - Clones, Linked Clones, VM Snaps
  - Writeable Snapshots (FlexClone, XtremIO, LVM Snaps)
- Building is harder than buying

Patterns?

Example Workflows

- Provision Service
  - Snap Prod
  - Spin up VM
  - Sanitize Data
- Push App
- Test
- Dispose

- Provision Service
  - Call CDM to provision
  - Sanitize
- Push App
- Test
- Dispose

AMI and Postgres Demo

- Provision Service
  - Snap Prod VM
  - Spin up VM
  - Allocate IP
  - Sanitize Data in PG
- Push App
- Test
- Dispose

https://github.com/krujos/data-lifecycle-service-broker


MAY 11-12, 2015 | SANTA CLARA, CA
CLOUD FOUNDRY
SUMMIT