

Don't Panic!

How to Cope Now You're
Responsible for Production

Euan Finlay
@efinlay24





Ahmen Khawaja @AhmenKhawaja · 2 hrs

False Alarm: Have deleted previous tweets!!



Ahmen Khawaja @AhmenKhawaja · 2 hrs

False Alarm to Queen's death! She is being treated at King Edward 7th hospital. Statement due shortly



Ahmen Khawaja @AhmenKhawaja · 2 hrs

"Queen Elizabeth has died": @BBCWorld



Ahmen Khawaja @AhmenKhawaja · 2 hrs

BREAKING: Queen Elizabeth is being treated at King Edward 7th Hospital in London. Statement due shortly:

[@BBCWorld](#)



December 3, 2015 12:36 pm

ECB leaves rates unchanged in surprising decision

Claire Jones, Frankfurt

Share

Author alerts

Print

Clip

Comments



The European Central Bank has left interest rates unchanged, dashing expectations of a cut to its deposit rate.



Financial Times ✓

@FinancialTimes



Follow

ECB leaves rates unchanged in shock decision on.ft.com/1Nrekqz

RETWEETS

36

LIKES

6



7:38 AM - 3 Dec 2015







On Thursday we published an incorrect story on FT.com that stated the European Central Bank had confounded expectations by deciding to hold interest rates rather than cut them. The story was published a few minutes before the decision to cut rates was announced.

The story was wrong and should not have been published. The article was one of two pre-written stories — covering different possible decisions — which had been prepared in advance of the announcement. Due to an editing error it was published when it should not have been. Automated feeds meant that the initial error was compounded by being simultaneously published on Twitter.

The FT deeply regrets this serious mistake and will immediately be reviewing its publication and workflow processes to ensure such an error cannot happen again. We apologise to all our readers.

OBITUARIES

Commander James Bond

Royal Navy and British Secret Service

Secret service agent James Bond
collaborator Wei Lin of the
People's External Security Force
died this morning in



/usr/bin/whoami

`/usr/bin/whodoiworkfor`

No such file or directory.

Trump Russia ties

US waits on findings of Mueller report

Justice department says no more indictments to come from special counsel



- The political battle continues after Robert Mueller's report
- Attorney-general's letter to Congress on Mueller report
- Mueller investigation: what we know, and what we don't

Trump Russia ties

Trump faces legal peril beyond Mueller probe

Prosecutors in New York's southern district are poring over the president's affairs

2 HOURS AGO



Brexit

Lidington disavows interest in replacing May

PM to meet Eurosceptic critics at Chequers after Tory MPs push for her to quit

AN HOUR AGO



Deutsche Bank AG

Deutsche staff banned from selling shares during merger

Top bankers told deferred stock is on 'no-sale' list until Commerzbank talks resolved



Germany

Berlin moves to cut ties between German Muslims and Turkey

Ankara's role under pressure as political tension rises after refugee crisis



Huawei Technologies Co Ltd

Huawei ramps up university investment despite US pressure

Chinese telecoms group reveals it spends \$300m a year on academic partnerships

Your team is now on call.

And you're mildly terrified.

Obligatory audience interaction.

**Everyone feels the same
when they start out.**

I still do today.

How do you get comfortable with supporting production?



The Ghosts of Incidents...

> Future

The Ghosts of Incidents...

Future

> Present

The Ghosts of Incidents...

Future

Present

> Past

A scene from the movie 'The Ghost of Christmas Future'. On the left, the Ghost of Christmas Future is depicted as a large, shrouded figure made of tattered, grey fabric, with a dark, hollow face. On the right, Ebenezer Scrooge is shown wearing a brown, patterned coat and a white, knitted hat with a tassel. He has a serious expression. The background is a dark, stone building with arched windows.

The Ghost of Incidents Future

**Handling incidents is the
same as any other skill.**

Get comfortable with your alerts.

**Delete the alerts
you don't care about.**

Have a plan for when things break.

Keep the documentation up to date.

Practice regularly.

***"The Gang Deletes
Production"***

NO PRODUCTION SERVERS?

NO PRODUCTION INCIDENTS

Opening
Mon
Tue-Thu
Fri-Sat
Sunday

Break things, and see what happens.

Did your systems do what you expected?



The Planned Datacenter Disconnect

We got complacent, and stopped running datacenter failure tests...

**Have a central place for reporting
changes and problems.**



19:25
Seeing aws dx link issues again-checking

Pasted image at 2018-07-27, 5:28 PM ▾



19:39
Methode alerts are firing

intermittent



19:40
yep we have network issues again at PR (edited)



Looks like the MPLS Verizon cct is down

So far no impact reported...

monitoring for now



19:44
We have reports of publishing not working, and problems with Methode portalpub connecting to UPP again



19:45
thanks

^^Verizon are saying PR site is affected by an issue affecting multiple locations



19:46
switching portalpub off in PR

We're not perfect.

But we always try to improve.

The Ghosts of Incidents...

Future

> Present

Past



The Ghost of Incidents Present

Calm down, and take a deep breath.

It's probably ok.

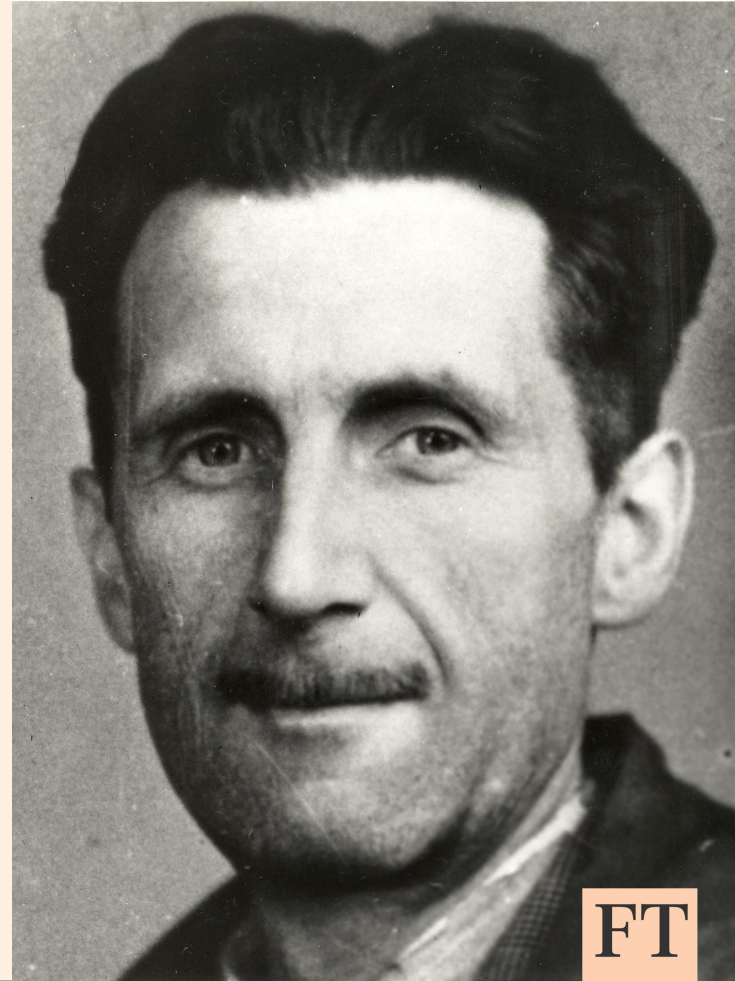
Don't dive straight in.

Go back to first principles.

What's the actual impact?

"All incidents are **equal**,
but some incidents are
more equal than others."

George Orwell, probably.



What's already been tried?



Is there definitely a problem?

DYNPUB / EXTRAPUB Prod Park Royal
Content

LAST UPDATED ON 10/21

DYNPUB / EXTRAPUB Prod Watford
Content

LAST UPDATED ON 10/21

SEMANTIC Prod Park Royal
Content

LAST UPDATED ON 10/21

SEMANTIC Prod Watford
Content

LAST UPDATED ON 10/21

SEMANTIC Prod AWS
Content
DownTime: 5

LAST UPDATED ON 10/21

DON'T
PANIC

OS/CO Prod US

OK
Healthy

MASHERV2-TESTS-PROD-UK



LAST UPDATED ON 10/21

MASHERV2-TESTS-PROD-US



LAST UPDATED ON 10/21

DON'T
WORRY

EVERYTHING
IS
FINE

PROBABLY
OK

SEMANTIC Test Park Royal
Content: 0

LAST UPDATED ON 10/21

SEMANTIC Test Watford

LAST UPDATED ON 10/21

NOT
CRITICAL

PAT - TEST



LAST UPDATED ON 11/22

DYNPUB / EXTRAPUB Int Park Royal
Content

LAST UPDATED ON 11/22

SEMANTIC Int Park Royal

LAST UPDATED ON 11/21

PAT - TEST



LAST UPDATED ON 11/22

What's the minimum viable solution?

Get it running *before* you get it fixed.

Go back to basics.

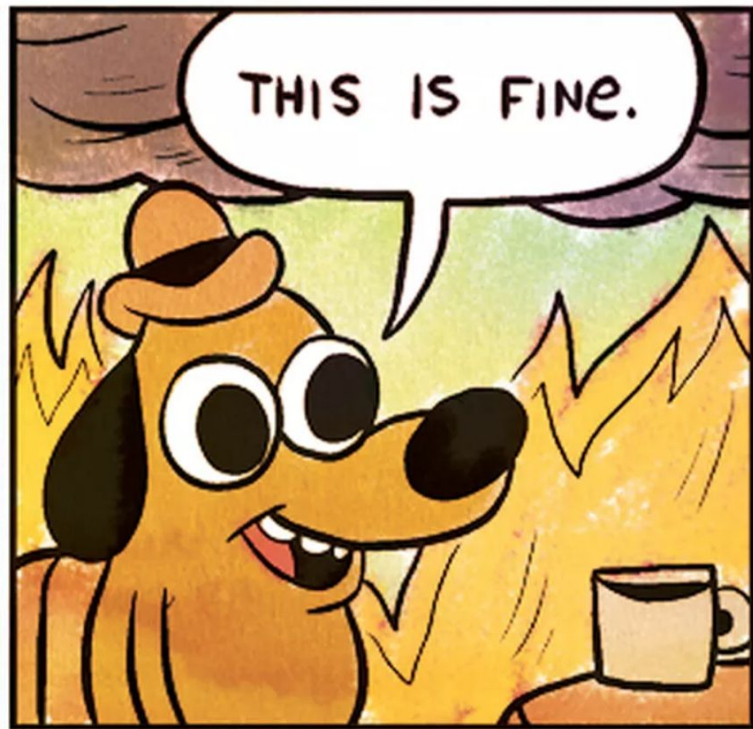
It's not DNS

There's no way it's DNS

It was DNS

-SSBroski





Don't be afraid to call for help.





**The One Where a Director
Falls Through the Ceiling**



Communication is key.

Especially to our customers.



Put someone in charge.



k8s - UPP Prod Delivery US: Annotations Read Aggregate Healthcheck is down
(Incident #4073269)

upp-prod-delivery-us.ft.com • [View details](#)

k8s - UPP Prod Delivery UK: Annotations Read Aggregate Healthcheck is down
(Incident #4073041)

upp-prod-delivery-eu.ft.com • [View details](#)

k8s - UPP Prod Delivery UK: Content Read Aggregate Healthcheck is down
(Incident #4073077)

upp-prod-delivery-eu.ft.com • [View details](#)

k8s - UPP Prod Delivery US: Content Publish Aggregate Healthcheck is down
(Incident #4073290)

upp-prod-delivery-us.ft.com • [View details](#)

k8s - UPP Prod Delivery US: Annotations Read Aggregate Healthcheck is up
(Incident #4073269)

upp-prod-delivery-us.ft.com • [View details](#)

k8s - UPP Prod Delivery US: Image Publish Aggregate Healthcheck is down
(Incident #4073407)

upp-prod-delivery-us.ft.com • [View details](#)

k8s - UPP Prod Delivery UK: Image Publish Aggregate Healthcheck is down
(Incident #4073095)

upp-prod-delivery-eu.ft.com • [View details](#)

Software can be chaotic, but we make it work



Expert

Trying Stuff Until it Works

○ RLY?

The Practical Developer
@ThePracticalDev

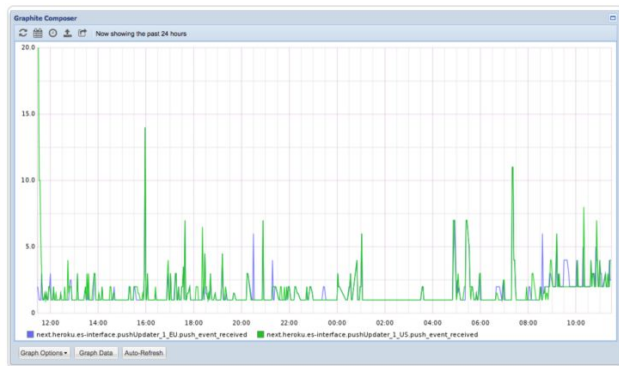
Create a temporary incident channel.

11:26
Same for `EA460F68-495E-D18F-3130-C220FA23EF8E` in prod-us at 04:01:57

11:26
So [redacted] says that in the last 24 hours, Next have got 20% more push event notifications in the US than in the EU...

11:26
Both request times line up exactly with the notification timestamps

11:27
uploaded and commented on this image: [Push notifications received](#)



“ green is US region

11:28
`suggestions-rw-neo4j` (that's v2 annotations as I understand it) wrote annotations at 04:01:36 (uk) and 04:01:51 (us)

So they would have been present before the notifications were sent out

11:29
[redacted] and I have confirmed that no annotations will still allow the article to show in both enrichedcontent and internalcontent

🙄 1

Hmm, that's a big gap between 1 and 5

**If you think you're
over-communicating,
it's probably just the right amount.**

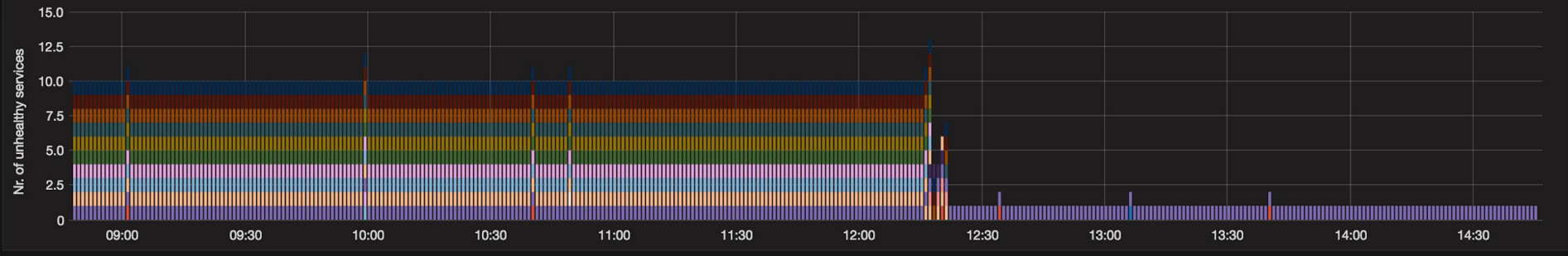
Tired people don't think good.



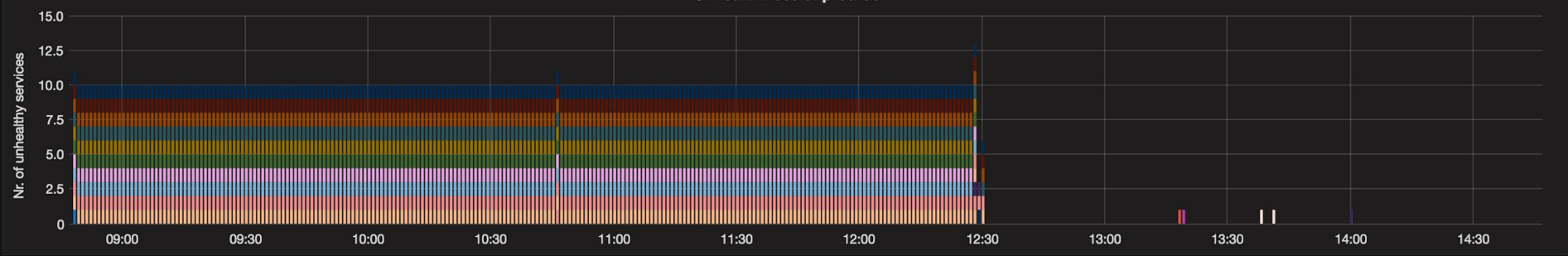
environment: prod-uk + prod-us + pub-prod-uk

services: All

Unhealthiness of prod-uk



Unhealthiness of prod-us



*"The Gang Serves Traffic
From Staging"*



Response time

Downtime

1h51m

Outages

21

Uptime

96.15%

Max resp. time

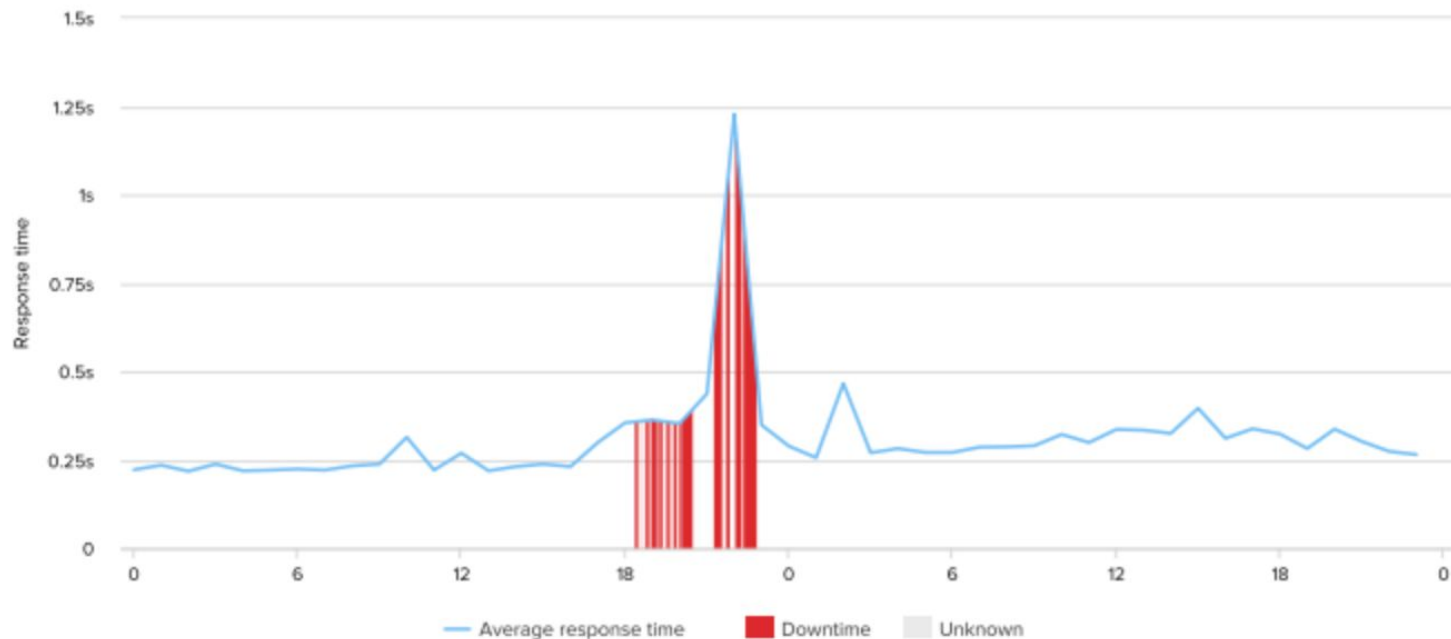
1.23s

Min resp. time

218ms

Avg resp. time

302ms



**It wasn't great,
but it wasn't the end of the world.**

The Ghosts of Incidents...

Future

Present

> Past

A man in a brown coat and white cap looks down with a somber expression. Behind him, a ghostly figure in white, possibly a woman, is visible. The scene is set in a dark, stone-walled environment.

The Ghost of Incidents Past

Congratulations! We survived.

It probably wasn't **that** bad, was it?

**Run a learning review
with everyone involved.**

Incident reports are important.

NEVER HAVE I FELT SO
CLOSE TO ANOTHER SOUL
AND YET SO HELPLESSLY ALONE
AS WHEN I GOOGLE AN ERROR
AND THERE'S ONE RESULT
A THREAD BY SOMEONE
WITH THE SAME PROBLEM
AND NO ANSWER
LAST POSTED TO IN 2003



Postmortem of database outage of January 31

Postmortem on the database outage of January 31 2017 with the lessons we learned.

[← Back to company](#)

On January 31st 2017, we experienced a major service outage for one of our products, the online service GitLab.com. The outage was caused by an accidental removal of data from our primary database server.

This incident caused the GitLab.com service to be unavailable for many hours. We also lost some production data that we were eventually unable to recover. Specifically, we lost modifications to database data such as projects, comments, user accounts, issues and snippets, that took place between 17:20 and 00:00 UTC on January 31. Our best estimate is that it affected roughly 5,000 projects, 5,000 comments and 700 new user accounts. Code repositories or wikis hosted on GitLab.com were unavailable during the outage, but were not affected by the data loss. GitLab Enterprise customers, GitHub customers, and self-hosted GitLab CE users were not affected by the outage, or the data loss.



"Until a restore is attempted, a backup is both **successful** and **unsuccessful.**"

Erwin Schrödinger?

Timeline

On January 31st an engineer started setting up multiple PostgreSQL servers in our staging environment. The plan was to try out [pgpool-II](#) to see if it would reduce the load on our database by load balancing queries between the available hosts. Here is the issue for that plan: [infrastructure#259](#).

± **17:20 UTC:** prior to starting this work, our engineer took an LVM snapshot of the production database and loaded this into the staging environment. This was necessary to ensure the staging database was up to date, allowing for more accurate load testing. This procedure normally happens automatically once every 24 hours (at 01:00 UTC), but they wanted a more up to date copy of the database.

± **19:00 UTC:** GitLab.com starts experiencing an increase in database load due to what we suspect was spam. In the week leading up to this event GitLab.com had been experiencing similar problems, but not this severe. One of the problems this load caused was that many users were not able to post comments on issues and merge requests. Getting the load under control took several hours.

We would later find out that part of the load was caused by a background job trying to remove a GitLab employee and their associated data. This was the result of their account being flagged for abuse and accidentally scheduled for removal. More information regarding this particular problem can be found in the issue "[Removal of users by spam should not hard delete](#)".

Publication of the outage

In the spirit of transparency we kept track of progress and notes in a [publicly visible Google document](#). We also streamed the recovery procedure on YouTube, with a peak viewer count of around 5000 (resulting in the stream being the #2 live stream on YouTube for several hours). The stream was used to give our users live updates about the recovery procedure. Finally we used Twitter (<https://twitter.com/gitlabstatus>) to inform those that might not be watching the stream.

The document in question was initially private to GitLab employees and contained name of the engineer who accidentally removed the data. While the name was added by the engineer themselves (and they had no problem with this being public), we will redact names in future cases as other engineers may not be comfortable with their name being published.

Data loss impact

Database data such as projects, issues, snippets, etc. created between January 31st 17:20 UTC and 23:30 UTC has been lost. Git repositories and Wikis were not removed as they are stored separately.

It's hard to estimate how much data has been lost exactly, but we estimate we have lost at least 5000 projects, 5000 comments, and roughly 700 users. This only affected users of GitLab.com, self-hosted instances or GitHub instances were not affected.

**Identify what can be
improved for next time.**

Improving recovery procedures

We are currently working on fixing and improving our various recovery procedures. Work is split across the following issues:

1. Overview of status of all issues listed in this blog post (#1684)
2. Update PS1 across all hosts to more clearly differentiate between hosts and environments (#1094)
3. Prometheus monitoring for backups (#1095)
4. Set PostgreSQL's max_connections to a sane value (#1096)
5. Investigate Point in time recovery & continuous archiving for PostgreSQL (#1097)
6. Hourly LVM snapshots of the production databases (#1098)
7. Azure disk snapshots of production databases (#1099)
8. Move staging to the ARM environment (#1100)
9. Recover production replica(s) (#1101)
10. Automated testing of recovering PostgreSQL database backups (#1102)
11. Improve PostgreSQL replication documentation/runbooks (#1103)
12. Investigate pgbarman for creating PostgreSQL backups (#1105)
13. Investigate using WAL-E as a means of Database Backup and Realtime Replication (#494)
14. Build Streaming Database Restore
15. Assign an owner for data durability

Nearly the end.

Don't clap yet.

Failure is inevitable.

And that's ok.

The end.

"Please clap."
Jeb Bush, 2016

We're hiring!

<https://ft.com/dev/null/>

@efinlay24

euan.finlay@ft.com