Google

Can we make Data Integrity easier?

Adrian Ratnapala
Site Reliability Engineer

Adapted from slides by Kristina Bennet and Raymond Blum.

Hope is not a strategy

- Reliance on data continues to grow in every dimension
  - Greater volumes of information
  - Growing numbers of platforms and uses
- The importance of preserving that information grows as well
- The way we preserve data has to evolve as fast as the data
  - Challenges of scale, both digital and physical
  - Generations of platforms and storage hardware

The Importance of Integrity

Availability: The data can be read on demand

At 99.99%, that's a total of less than 53 minutes of downtime in a year.

Generally, users recover, and life goes on.

Integrity: The data is exactly what was written

Imagine 99.99%, for 2 GiB, that would be 200+ KiB corrupted or missing.

Consequences are often long lasting, and possibly unrecoverable.

- A document has lost several pages
- An executable is useless
- A database is corrupt
- A video is garbled

And that's just for 2 GiB. *If you have 2 PiB, this might be happening to millions of users.*

What do we need?

Recovery

You need a trustworthy plan for how to **recover** from an outage. Depending on your system & priorities, you might:

- Restore from backups.
- Rebuild from an external source of truth
- Do both (e.g. for point-in-time restores).

A few words about replication

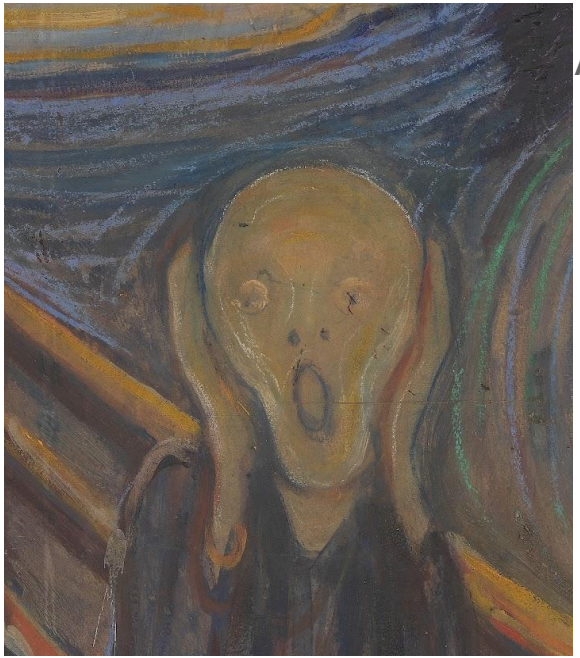
It's great but:

- Replication replicates data loss and corruption too.
- Same software stack. Same storage devices. Same vulnerabilities.

But it's not enough to just fish data out of cold storage.

Recovery means serving correct data in production.

You need recovery plan



Are you sure that:

- You really did back the right data up.
- You can find backup.
- How you can recover the data back into service.
- What exactly the recovery process will **do.**
- That you have the required storage and processing capacity.
- Recovery won't break or corrupt anything.
- You can recover it quickly enough.

How will you know all this **before** the bad thing happens?

What do I do with a plan?

Practice! practice! practice!

- **Prediction:** your recovery will fail on the first try.
- Failure is no surprise, but, the *way* it fails might be.

Automate! Automate! Automate!

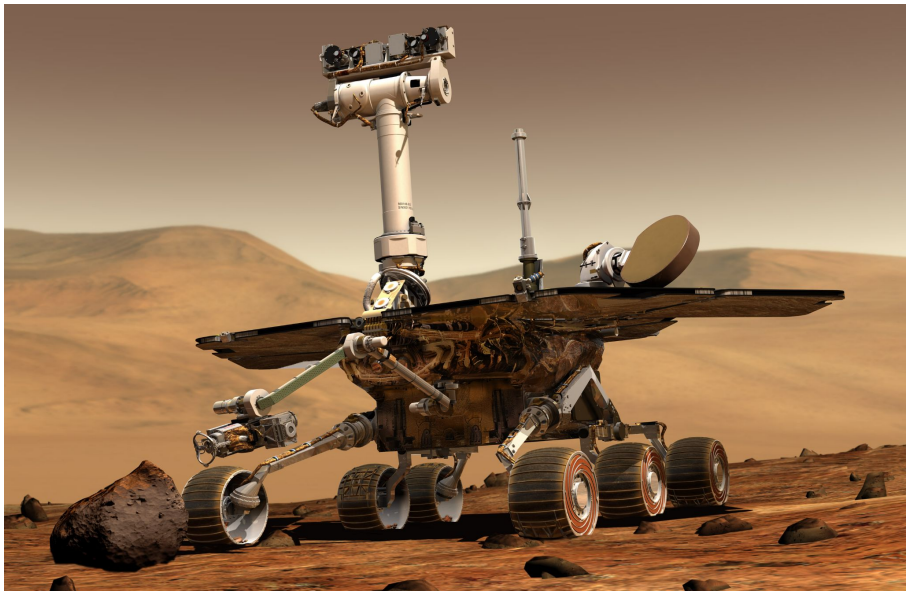
- No need to relearn the process each time.
- No trying to decipher incomplete and out-of-date notes
- Anybody can run the automation, not just the one who quit two months ago.
- **No scrambling for answers when the crisis hits**

Test! Test! Test!

- Now you can run your automation any time. Run it all the time!
- Test often. Detect early. Adapt quickly.

Isn’t that hard work?

Repeating all that for every dataset owned by every team is **TOIL**.



Engineering a re-usable solution is an **OPPORTUNITY**.

The mission is …

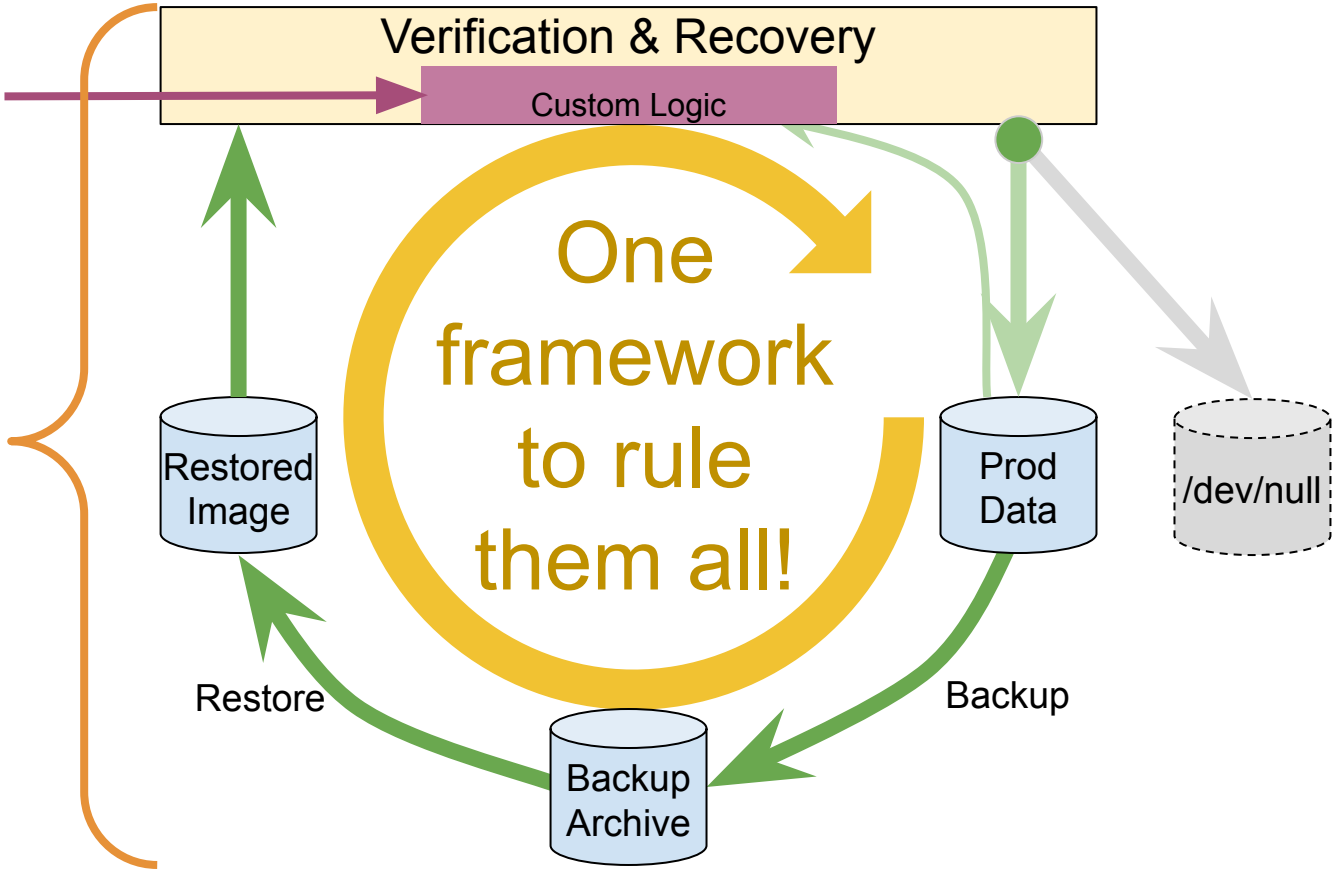
To build a **reusable, managed framework** to backup, restore & recover a datastore according to a **tested plan**.

So what might that look like?

The dream is

A few x100 lines of code & config

Drives all of this



So how do we get that?

Solve technical challenges

This is a big integration challenge, it must

- orchestrate lots of moving parts.
- sort out permissions
- sort out provisioning
- etc.

But we are SREs, we know how to do that stuff.

Face uncomfortable truths.

Toil isn't the only nasty truth of data recovery

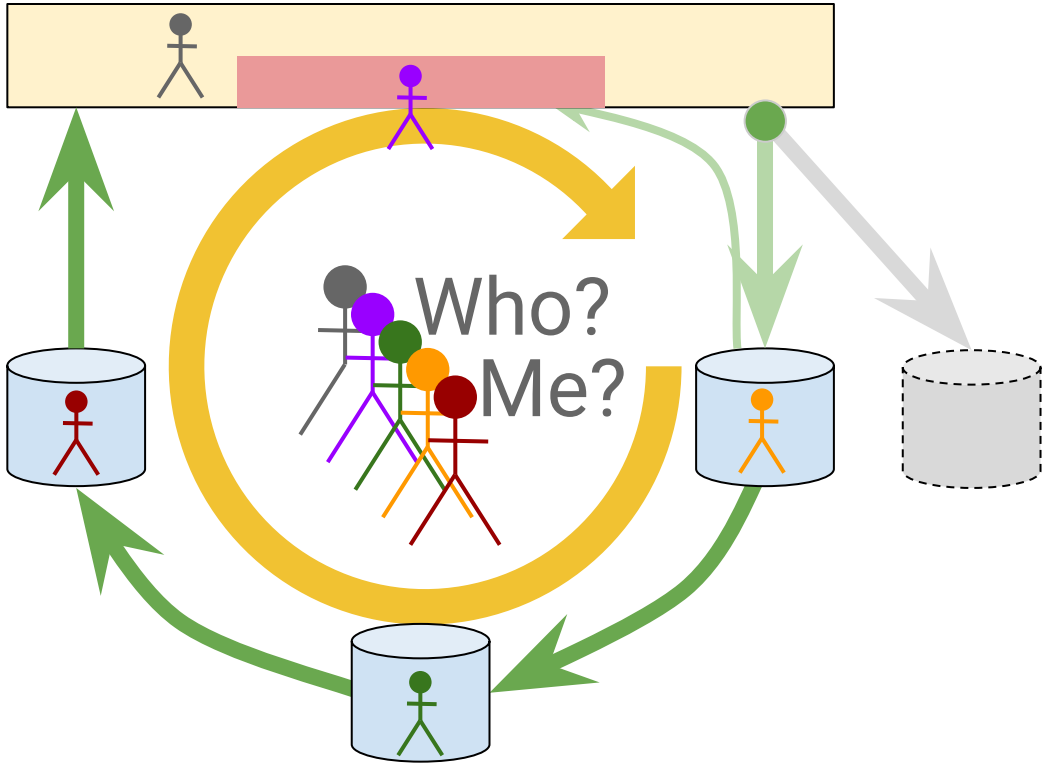
- Continuous testing → continuous cost.
- Writing back to prod is risky.
- Some databases will melt if you write lots of data.

Work across divisions in knowledge & responsibility

Lots of stakeholders hold pieces of the puzzle

- The database maker
- The application maker
- The backup solution maker

Whoever solves the puzzle, must communicate across those lines.



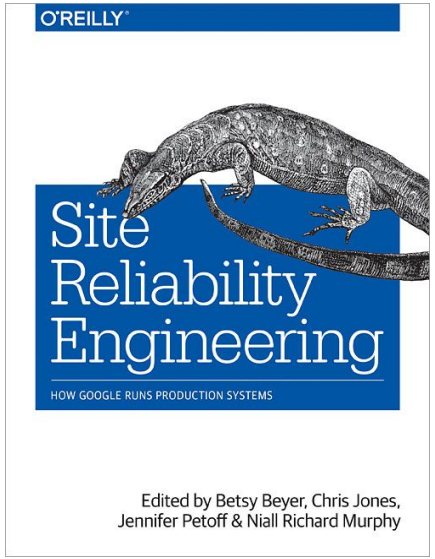
So let’s get started

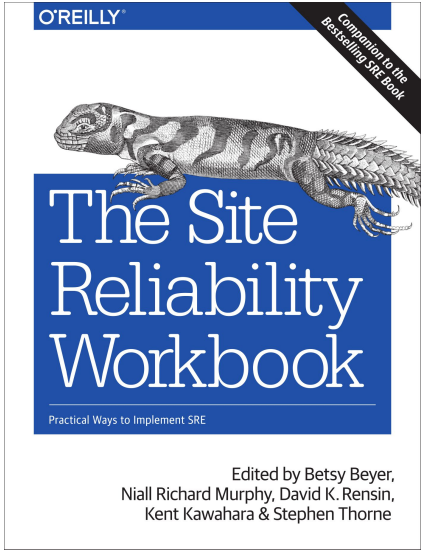
There is an opportunity to engineer a reusable framework to make data integrity much easier

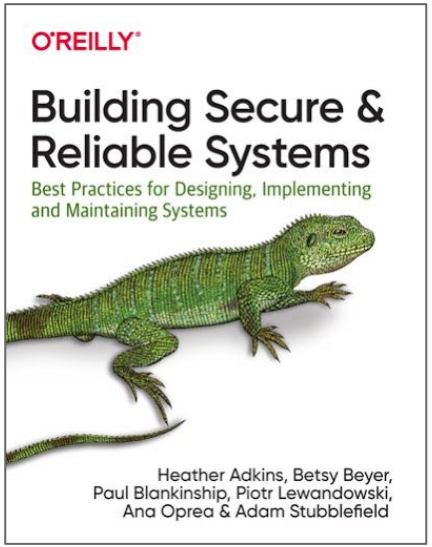
- People in this room are in a position to seize that opportunity.
- But to get there, we’ll need to talk to each other

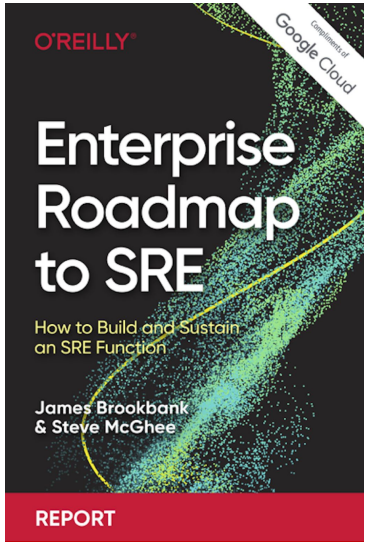
So let’s do that!

Find Google SRE publications—including the SRE Books, articles, trainings, and more—for free at sre.google/resources.









Google