Learning From Our Mistakes
Helpful or Evasive?

We’re currently experiencing degraded performance issues with [redacted]. Our team is currently working to restore normal performance levels. We apologize for any inconvenience. Users may be affected.

“We continue to investigate the issue affecting data services,” a [redacted] spokesman said. “Many of our services are starting to restore. We’re sorry for any issues caused and will provide an update when we know more.”

The company is still investigating the cause of the outage.

Aug 22, 2021

A subset of customers in Australia are unable to access [redacted].

Resolved - We received reports and resolved the issue causing a subset of customers in Australia unable to access [redacted].
At 15:56 PM PT, we received the reports of a subset of Australian customers unable to access [redacted].
At 16:18 PM PT, our engineering team resolved this issue.

Aug 22, 2021, 16:28 PDT
Why are these reports so evasive?

• Do the service providers see outage reports as an admission of some form of liability?
• Do they think that an open and direct appraisal of faults in their service will cause them reputational harm in the eyes of their customers?
• Or do they think that describing the causes of an outage somehow compounds their potential liabilities?
• What are they covering up here?
Helpful or Evasive?

Early today Facebook was down or unreachable for many of you for approximately 2.5 hours. This is the worst outage we’ve had in over four years, and we wanted to first of all apologize for it. We also wanted to provide much more technical detail on what happened and share one big lesson learned.

The key flaw that caused this outage to be so severe was an unfortunate handling of an error condition. An automated system for verifying configuration values ended up causing much more damage than it fixed.

The intent of the automated system is to check for configuration values that are invalid in the cache and replace them with updated values from the persistent store. This works well for a transient problem with the cache, but it doesn’t work when the persistent store is invalid.

Today we made a change to the persistent copy of a configuration value that was interpreted as invalid. This meant that every single client saw the invalid value and attempted to fix it. Because the fix involves making a query to a cluster of databases, that cluster was quickly overwhelmed by hundreds of thousands of queries a second.

To make matters worse, every time a client got an error attempting to query one of the databases it interpreted it as invalid, and deleted the corresponding cache key. This meant that even after the original problem had been fixed, the stream of queries continued. As long as the databases failed to service some of the requests, they were causing even more requests to themselves. We had entered a feedback loop that didn’t allow the databases to recover.

The way to stop the feedback cycle was quite painful - we had to stop all traffic to this database cluster, which meant turning off the site. Once the databases had recovered and the root cause had been fixed, we slowly allowed more people back onto the site.

This got the site back up and running today, and for now we’ve turned off the system that attempts to correct configuration values. We’re exploring new designs for this configuration system following design patterns of other systems at Facebook that deal more gracefully with feedback loops and transient spikes.
Helpful or Evasive?

Root Cause Analysis – What Happened

Contributory Factors – What exacerbated the incident?

Short Term – what we did to fix it

In Progress – what we doing right now

Long Term – how we intended to avoid repeating this outage
Why is this important?

• The internet is now the foundation for all of our communications - from the trivial and frivolous through to vital and life saving systems

• This is now a public safety issue, and we need to move away from the handling of operational incidents as PR exercises and take steps that other industries have already embraced
What are we doing about it?

• Time to call up Sean and Jared and have a chat