March 8th, 2018 – QCon London

How to build observable distributed systems

@PierreVincent  pvincent.io
Reaching production is only the beginning
No system is immune to failure
Be ready to recover
When distributing a system, we’re also distributing the places where things might go wrong.
Monitoring only applies to known failure modes. What about everything else?
“Monitoring tells you whether the system works. Observability lets you ask why it's not working.”

– Baron Schwarz
Healthchecks
Healthchecks

- Is it running?
- Can it perform its task?
- Can it accept more work?
Healthchecks

Broadcast

Register

Expose

@PierreVincent
GET http://1.2.3.4:8080/health

200 OK

{
    "service": "registration-service",
    "healthy": true,
    "workload": { "healthy": true },
    "dependencies": [
        { "name": "cassandra", "healthy": true },
        { "name": "billing-svc", "healthy": true }
    ]
}
Overzealous Healthchecks can be counter-productive

Source: HTTP Healthchecks for a Resilient Platform - Chris O’Dell
skeltonthatcher.com/blog/http-healthchecks-for-a-resilient-platform
Metrics
Prometheus

- Servers / VMs
  /metrics
- Appliances/Infra
  /metrics
- Services
  /metrics

Grafana

@PierreVincent
Watch out for over-reliance on metrics

- Not every metric deserves an alert
- Limitations at high-cardinality
- Real-time querying means some trade-offs on retention
- Limit alerting to user-impacting symptoms
- Poor fine-grained debugging e.g. CustomerId
- Not suitable for long-term trend analysis
Logging
Making sense of (a lot of) logs

Centralised

Searchable

Correlated
ERROR [svc=A][trace=a1b2c3] Failed to process order
  Cause: Order process manager responded with 500

ERROR [svc=F][trace=a1b2c3] Failed to complete order
  Cause: Shipping service responded with 500

INFO [svc=G][trace=a1b2c3] Items verified in stock

ERROR [svc=H][trace=a1b2c3] Failed to save order
  Cause: Cassandra timeout exception

@PierreVincent
2018-02-20T16:38:23+00:00  ERROR  Read timed out

When did it happen?
- timestamp: 2018-02-20T16:38:23+00:00

What is the message?
- level: ERROR
- log: Read timed out

What is it?
- service: registration-service
- team: events

What is it running?
- commit: 542a8b8e
- build: 542a8b8e.7
- runtime: java-1.8.0_161

Where is it?
- region: europe-west2
- node: node_e79f3e52

Who caused it?
- customerID: 55123
- userID: 458

Can I trace it?
- requestID: ec667cb45

Any other info?
- ...

Hmmm thanks... ?
Structured logs unleash high-cardinality exploration

Error rate spike isolated by build version

Activity spike isolated for single customer

@PierreVincent
Tracing
Tracing

get_confirmed_attendees

cassandra/select

get_attendees

cassandra/select

get_confirm_status

mysql/select

Spans

0ms  50ms  100ms  150ms  172ms

73ms

54ms

78ms

41ms

event-mgt-api

attendees-service

registration-service

@PierreVincent
Usability of tooling is key to adoption

Cheap to instrument

Easy to explore

Reliable & Trustworthy
Visibility builds trust but requires safety
Visibility helps justify decisions
Visibility enables operability
Don’t spend all your time testing...
...keep some for instrumenting.

Thank you!
Thank you!

Pierre Vincent
SRE Manager at Poppulo

@PierreVincent
pvincent.io