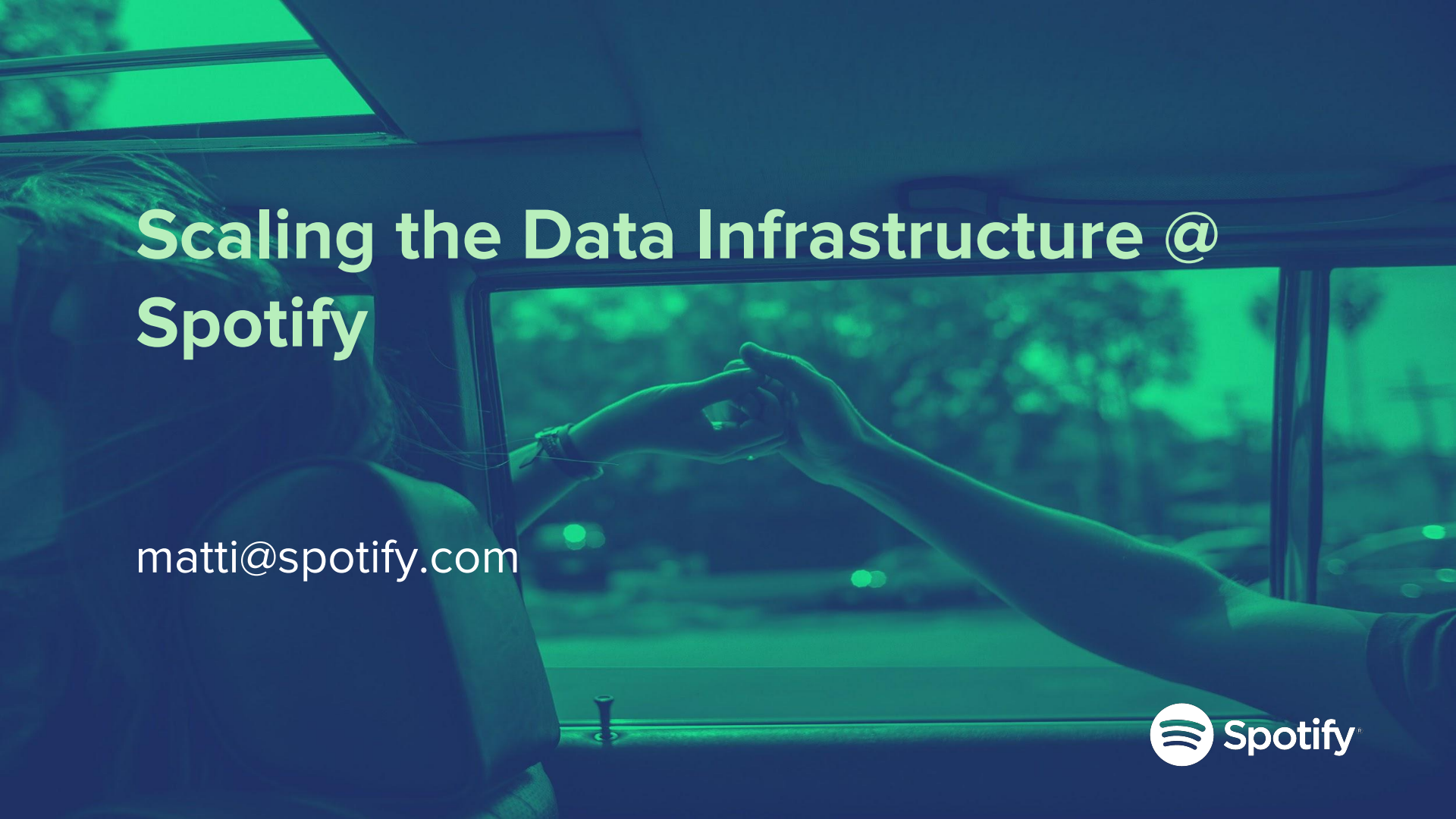


Scaling the Data Infrastructure @ Spotify

matti@spotify.com

Spotify

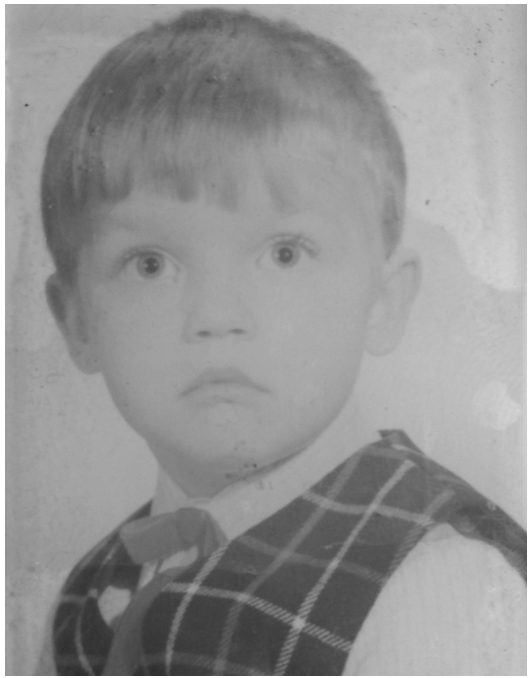
Matti Pehrs

> 25 years in IT

Emacs since 1987

Java since 1995

Spotify since 2013



Agenda

1. **Data at Spotify**
2. **Summer of 2015**
3. **Challenges & Victory**
    - Datamon
    - Styx
    - GABO

Data at Spotify

Hadoop

Data history through Hadoop

- ➔ Started with 5 servers in the office 2007
- ➔ Tried Amazon Elastic Map/Reduce 2010
- ➔ In 2012 we started using Hortonworks HDP
- ➔ We now run a 2000+ node cluster in London
- ➔ We are out of physical space!

Spotify is moving to the cloud

Spotify

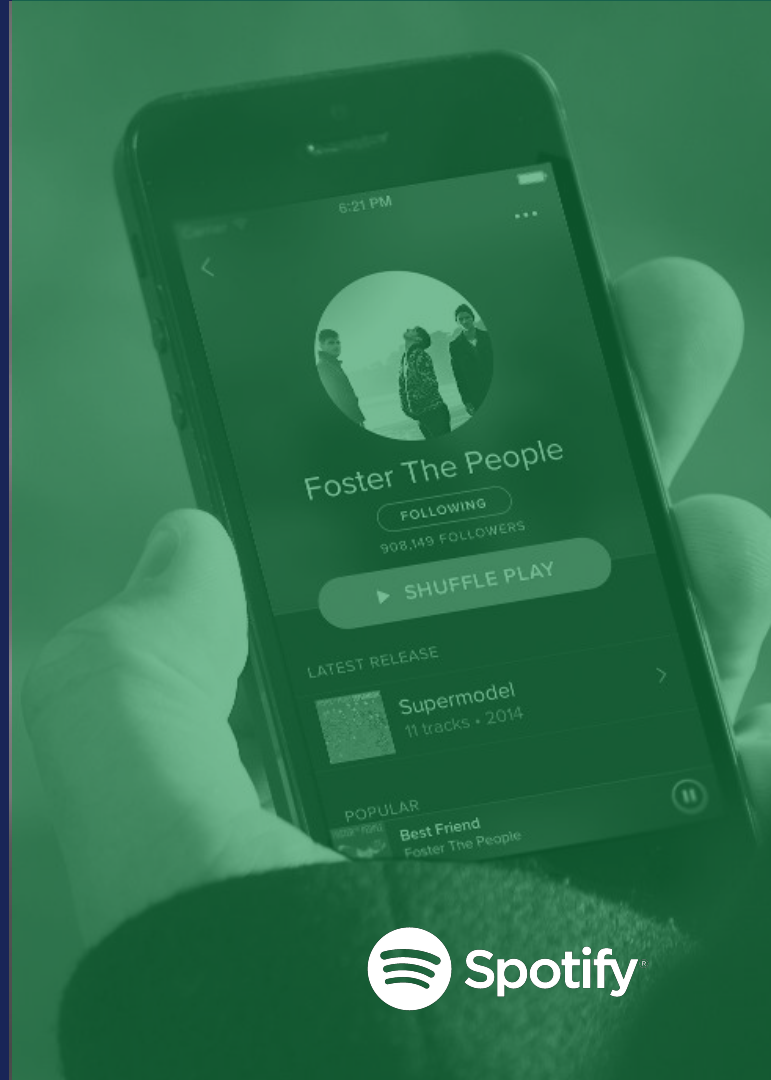
It’s all about focus.

Spotify’s core business is to serve music not operate data centers.

Spotify®

Spotify big-data context

- Over 100 million monthly active users
- Over 30 million song
- Over 2 billion playlists
- Active in 60 markets



Our growth in Data

Users

+50 TB/day

Developers

+60 TB/day
+10k M/R jobs

Data is at the heart of Spotify

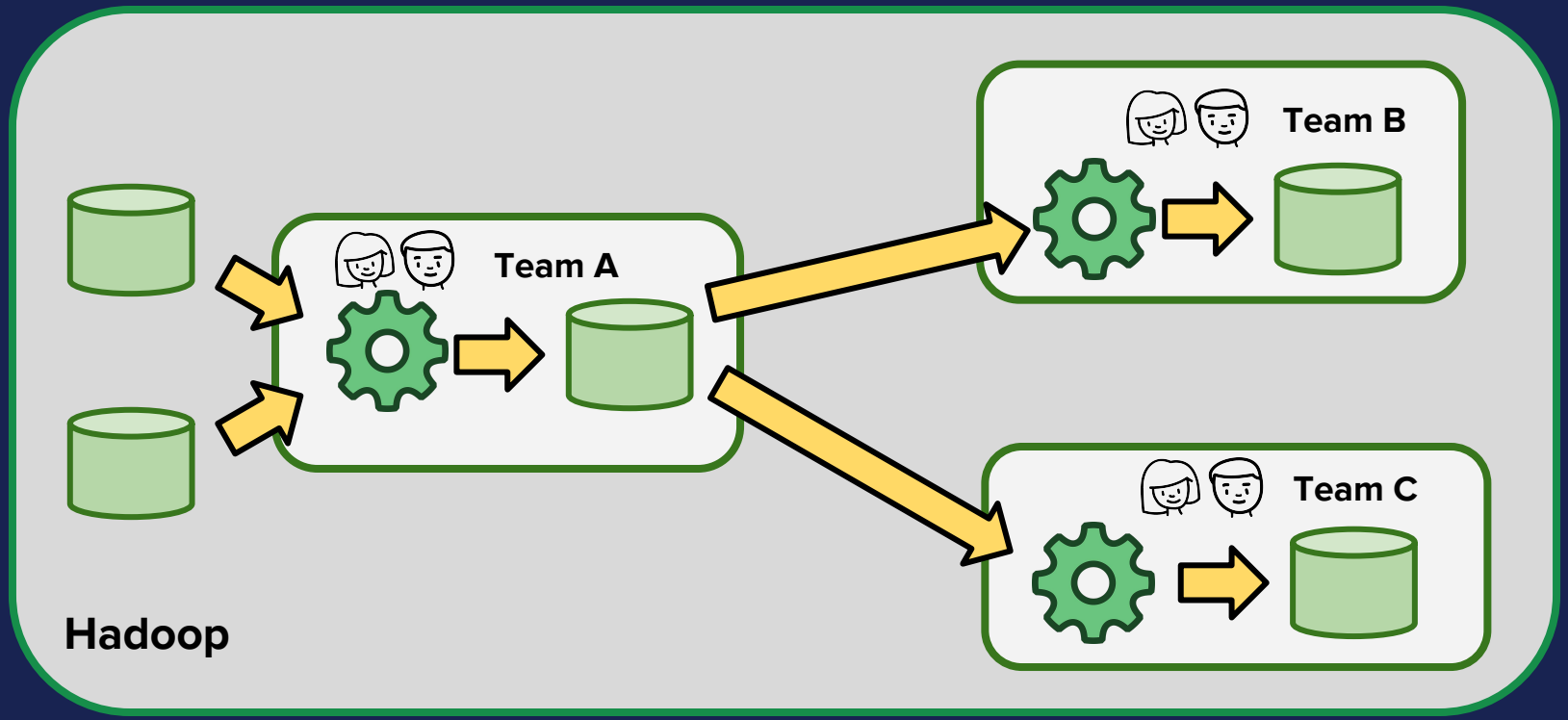
In 2007

- Reporting

In 2017

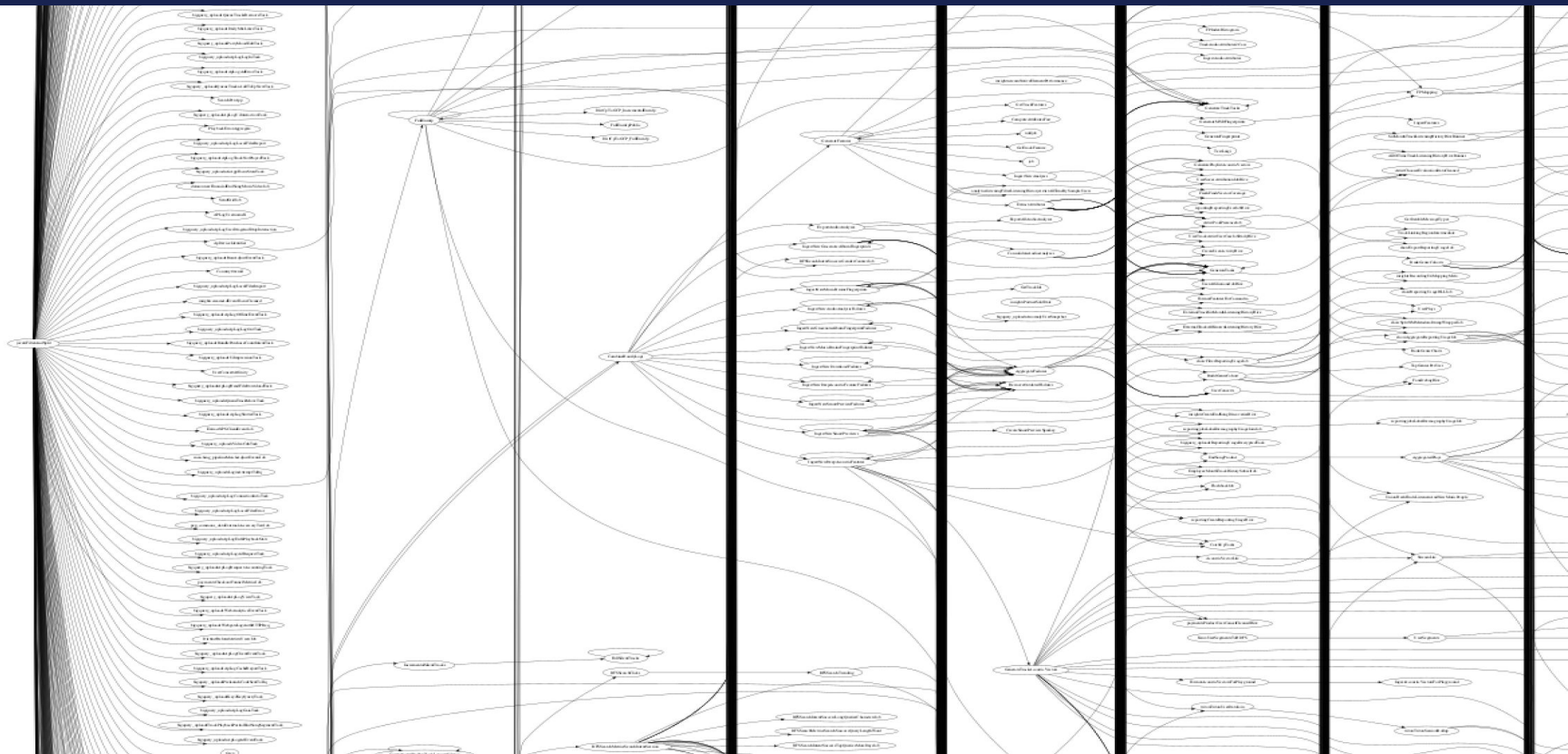
- Reporting
- All features use
  big data in some form

Autonomy & Dependencies

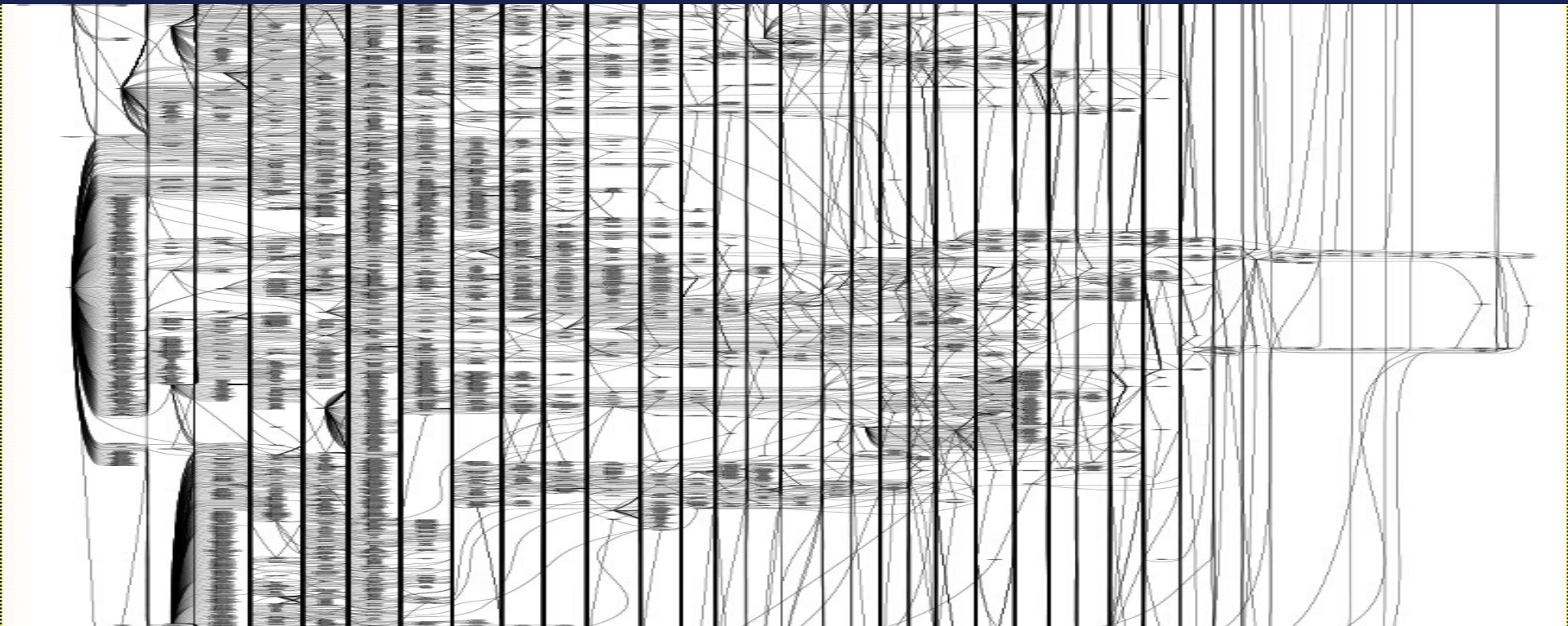


Autonomy & Dependencies

Autonomy & Dependencies

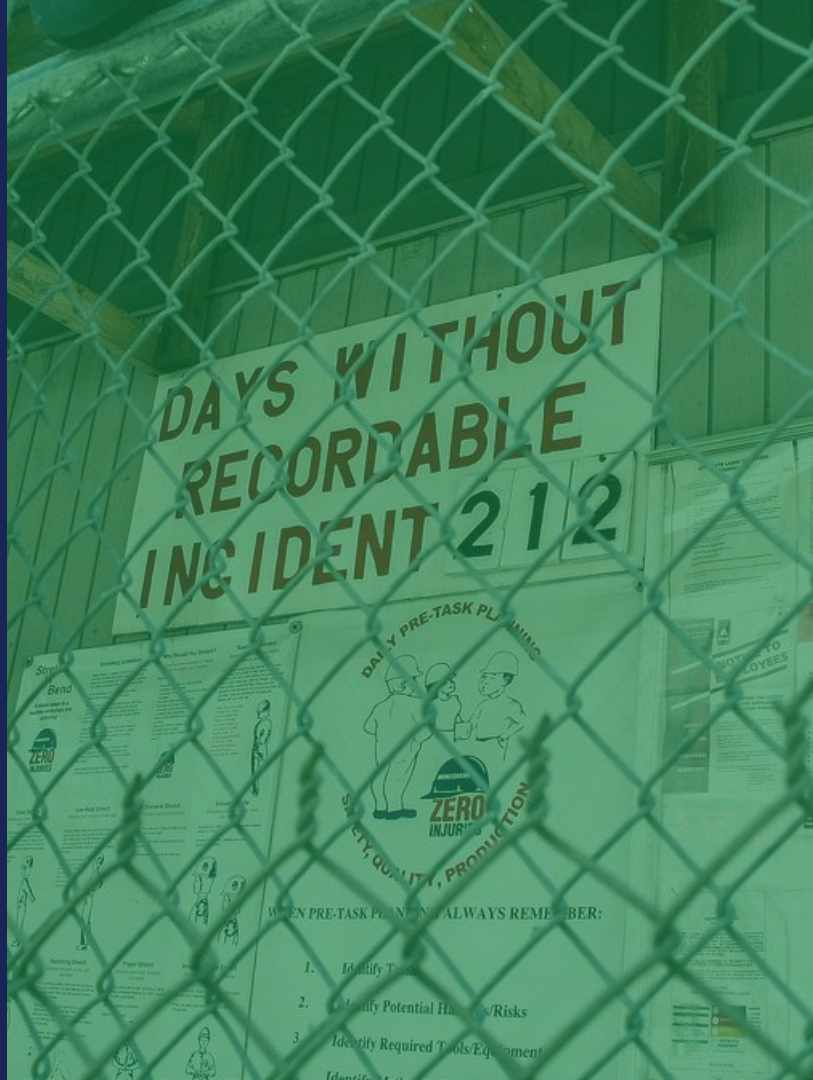


Autonomy & Dependencies



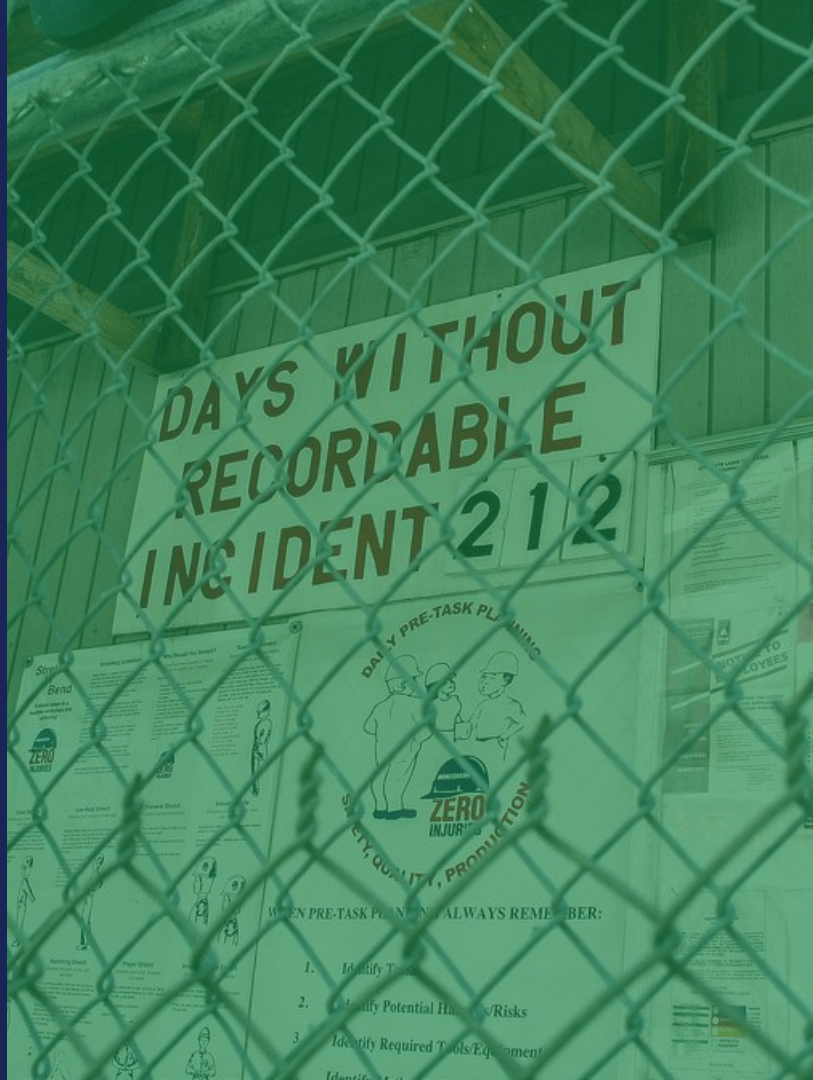
Summer of 2015

Summer of Incidents



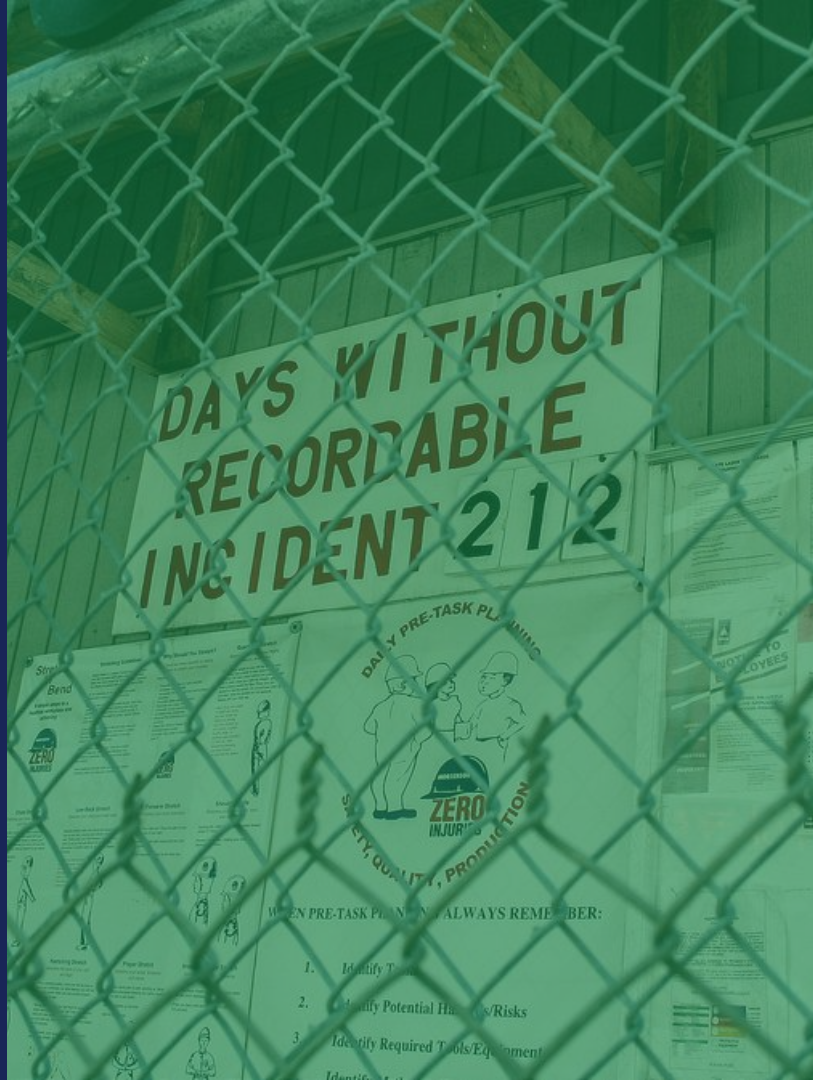
Summer of Incidents

- A strain of incidents



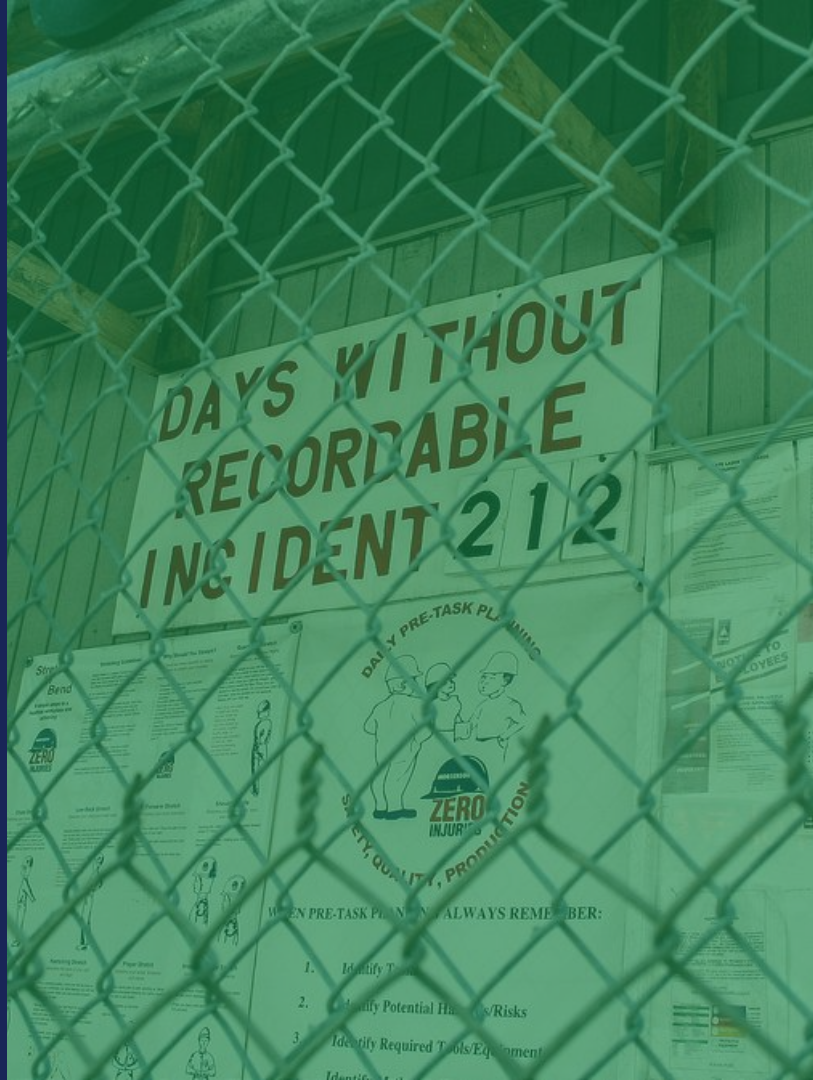
Summer of Incidents

- A strain of incidents
- War-room



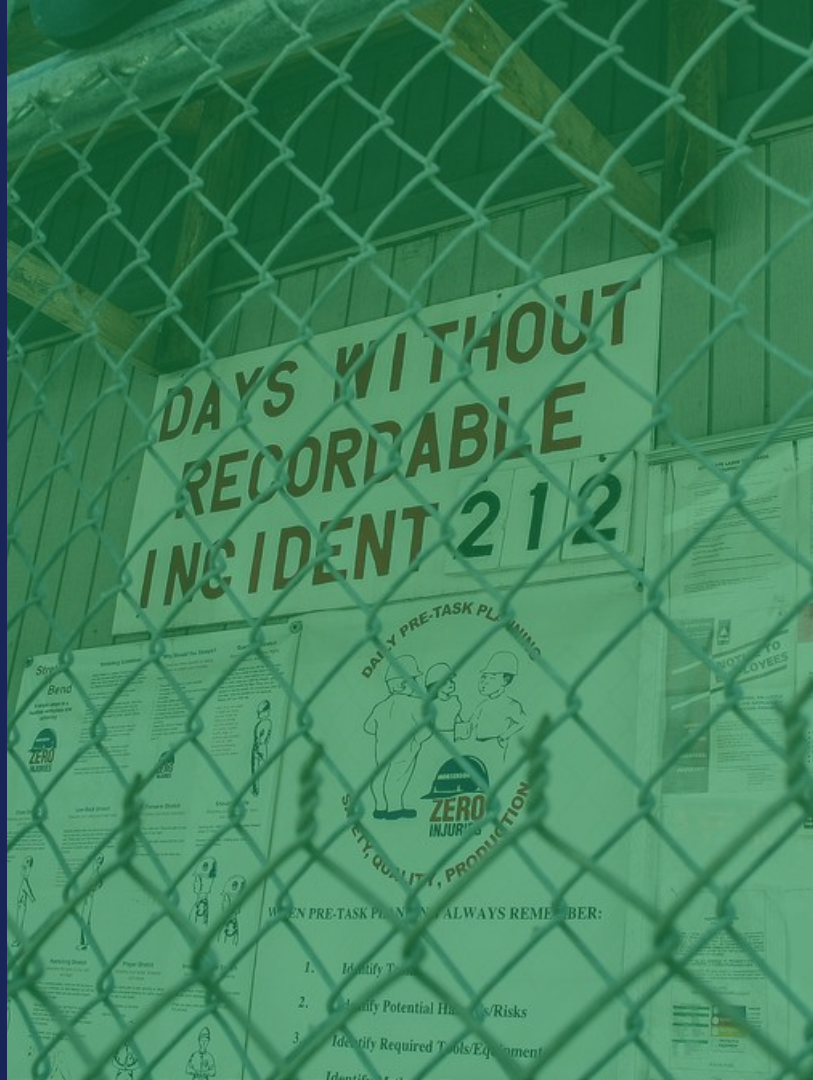
Summer of Incidents

- A strain of incidents
- War-room
- Hadoop on it’s knees



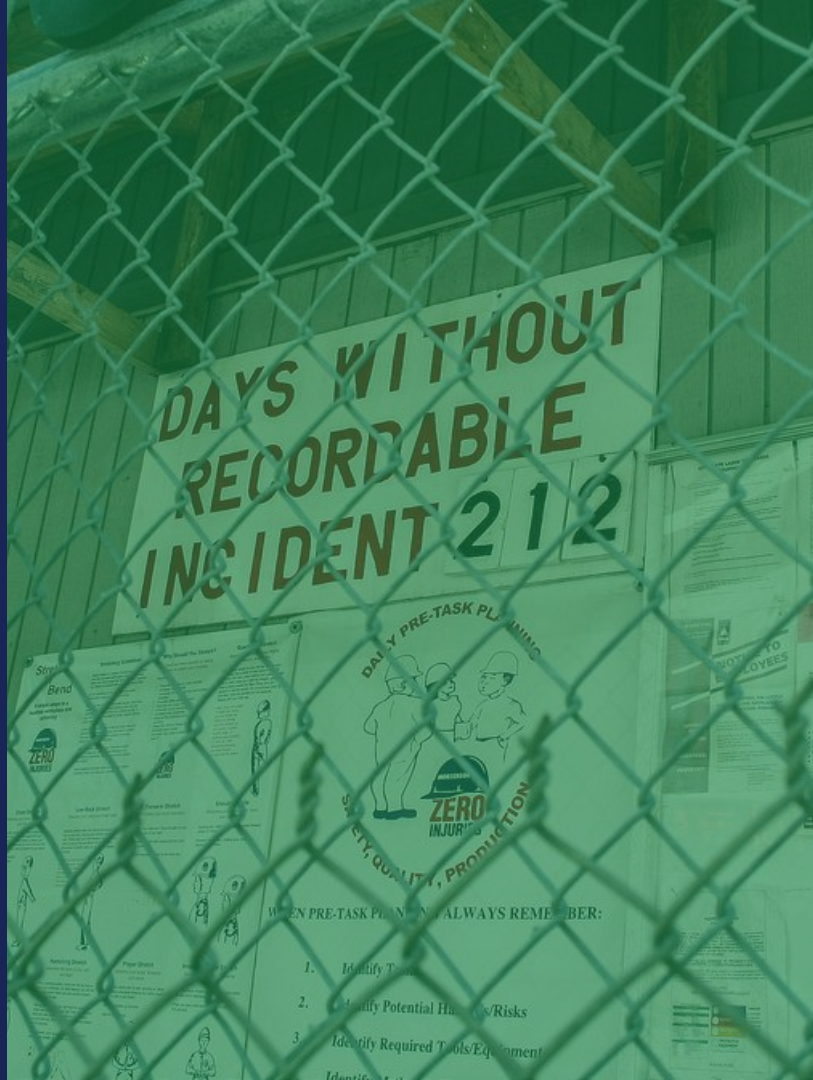
Summer of Incidents

- A strain of incidents
- War-room
- Hadoop on it’s knees
- Event Delivery Catch up



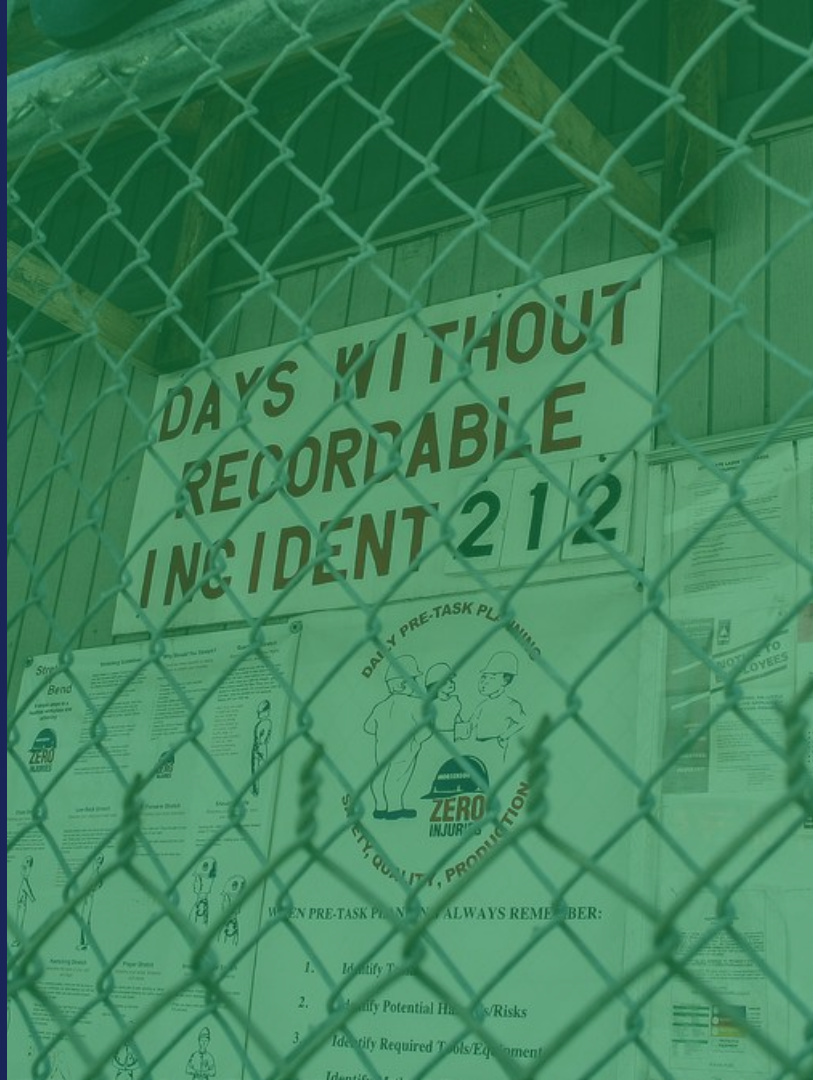
Summer of Incidents

- A strain of incidents
- War-room
- Hadoop on it’s knees
- Event Delivery Catch up
- Reprocessing of data



Summer of Incidents

- A strain of incidents
- War-room
- Hadoop on it’s knees
- Event Delivery Catch up
- Reprocessing of data
- Hard to debug data issues



Challenges and the path to victory...

Challenges and the path to victory...

1. Early Warning

Datamon - Data monitoring

Challenges and the path to victory...

1. Early Warning — Datamon - Data monitoring
2. Debuggability & Control — Styx - Scheduling and control

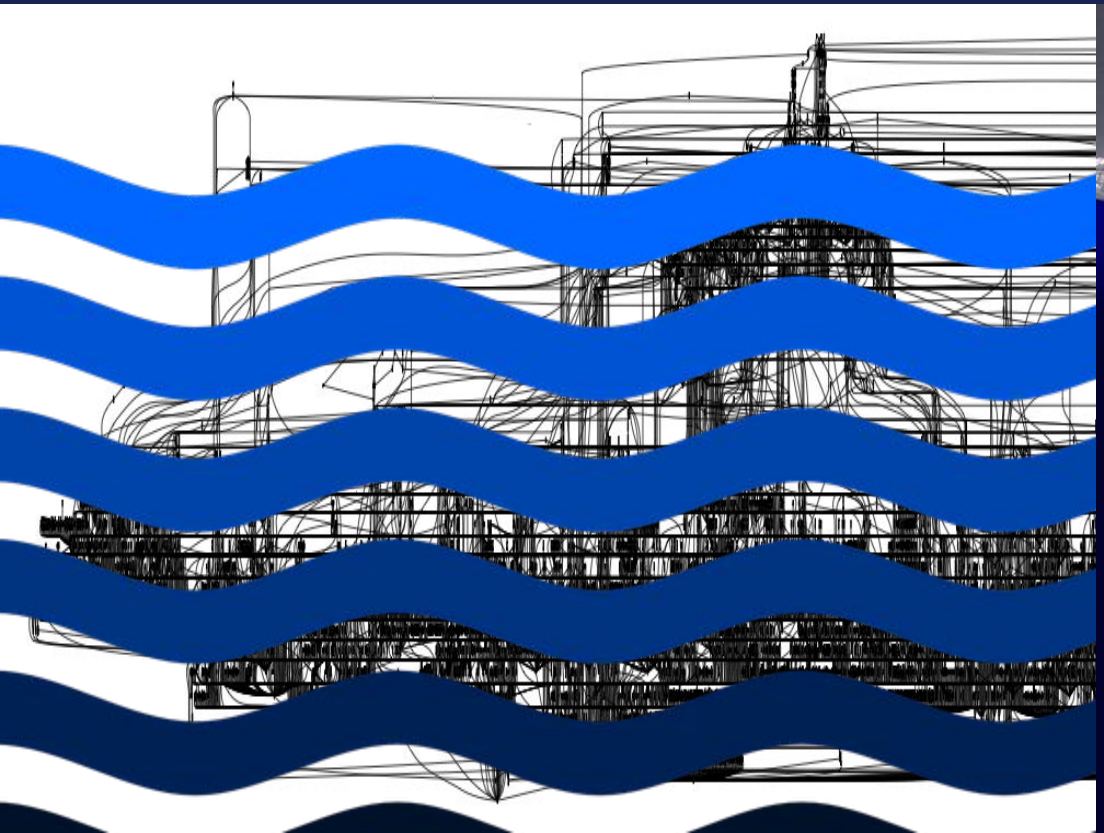
Challenges and the path to victory...

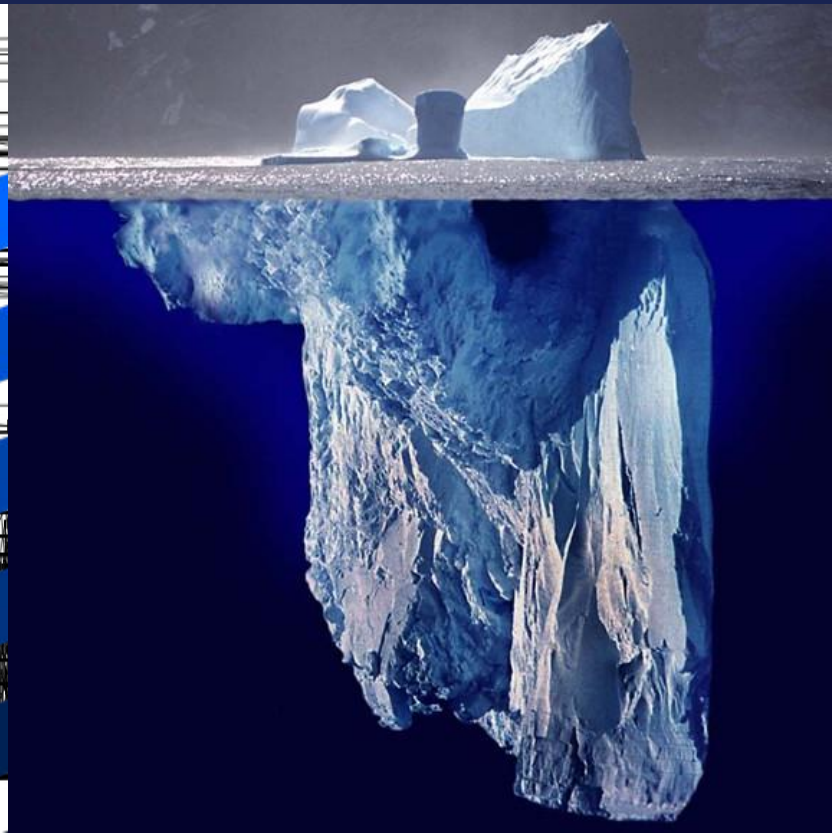
1. Early Warning — Datamon - Data monitoring
2. Debuggability & Control — Styx - Scheduling and control
3. Automate Capacity — GABO - Event Delivery

Challenges and the path to victory...

1. Early Warning — Datamon - Data monitoring
2. Debuggability & Control — Styx - Scheduling and control
3. Automate Capacity — GABO - Event Delivery

Early Warning - Datamon



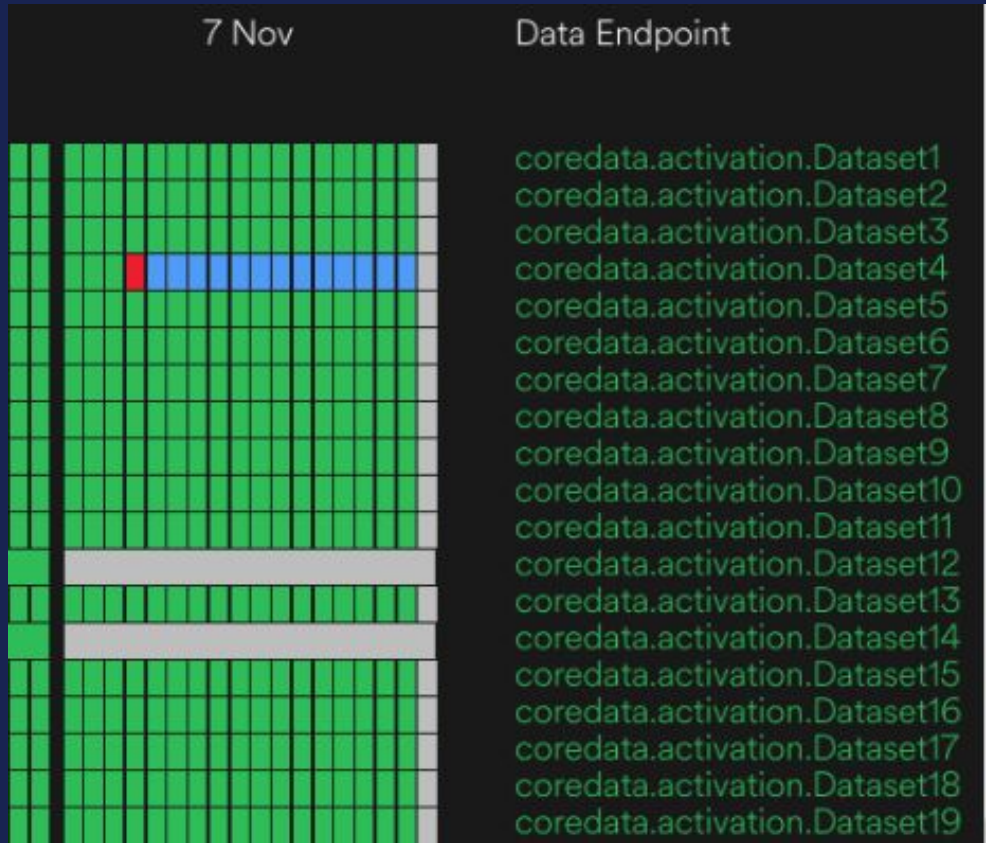


Early Warning - Datamon

- **Unified view**
  - Alignment between teams
- **Ownership**
  - Clear ownership of data
- **SLA**
  - Alert on late data

Early Warning - Datamon

- Define terminology
- Provide metadata language
- Implement a Datamon service



Challenges and the path to victory...

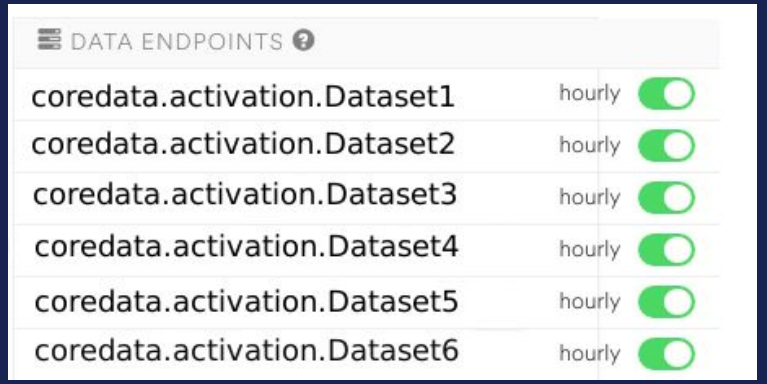
1. Early Warning — Datamon - Data monitoring
2. Debuggability & Control — Styx - Scheduling and control
3. Automate Capacity — GABO - Event Delivery

Debuggability & Control - Styx

- **Execution control**
  - Self service for data users
- **Execution information**
  - Expose debug information
- **Execution isolation**
  - Docker for data jobs

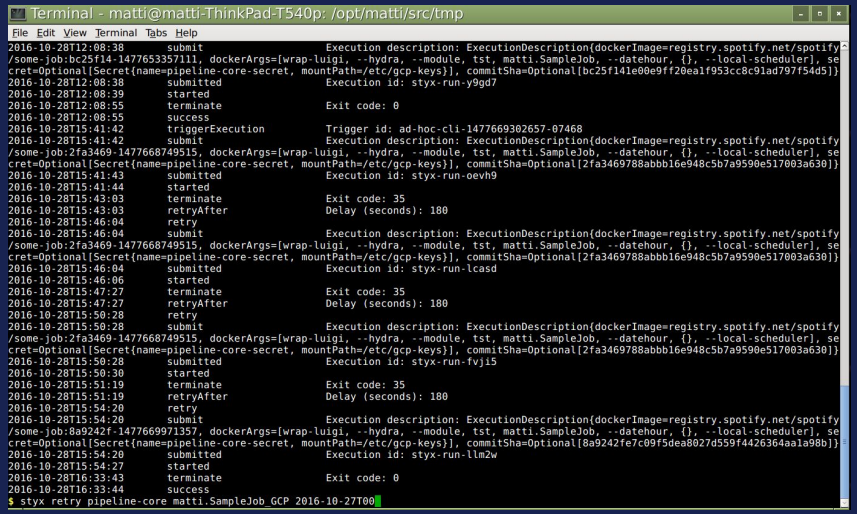
Debuggability & Control - Styx

- **Execution control**
  - Centralized execution API



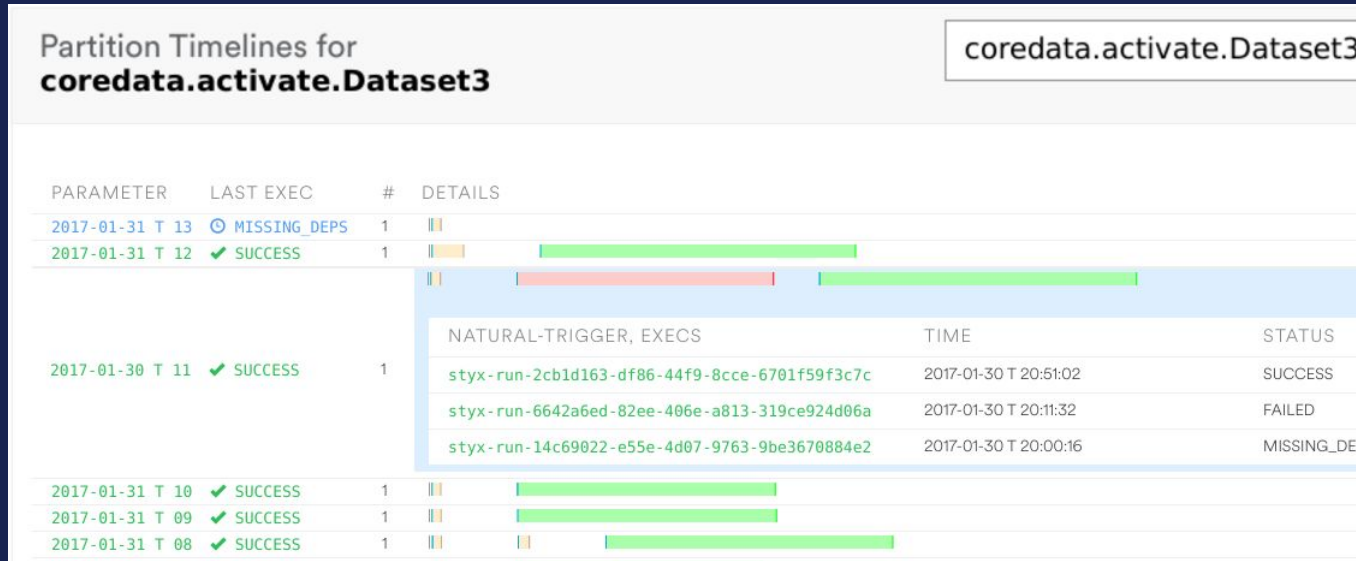
Debuggability & Control - Styx

- **Execution control**
  - Centralized execution API
  - Backfilling and reprocessing



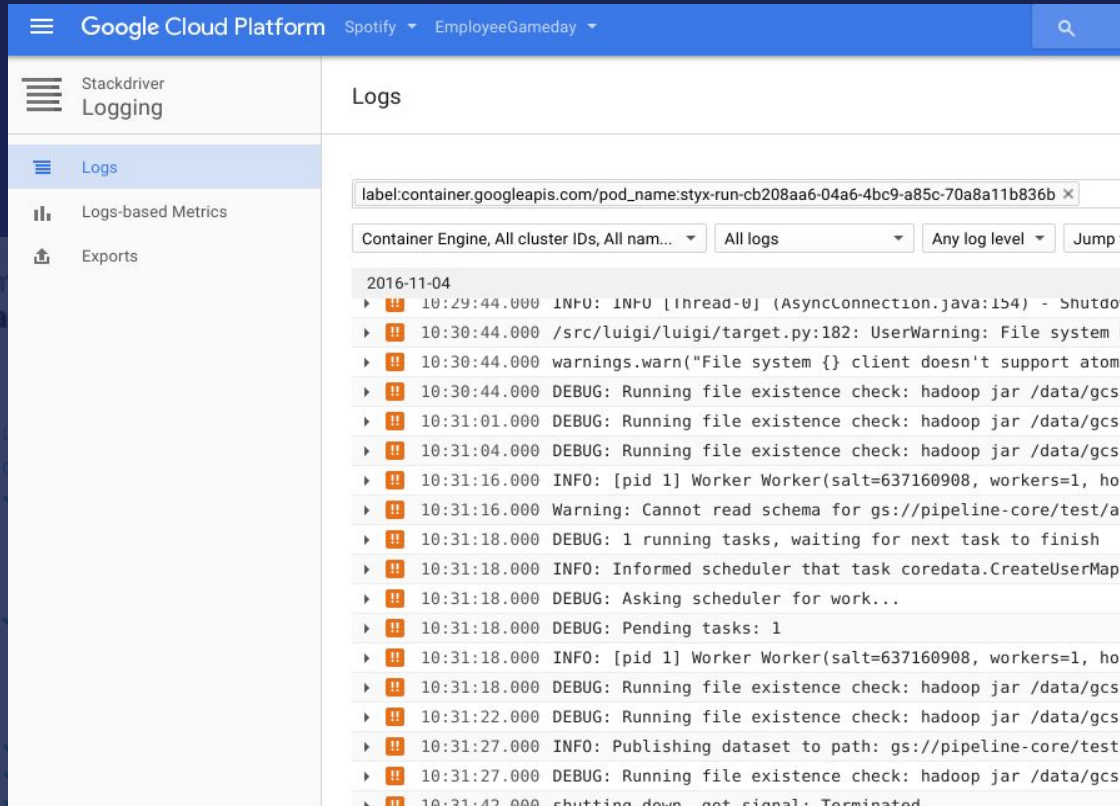
Debuggability & Control - Styx

- Execution control
- **Execution information**
  - Timeline



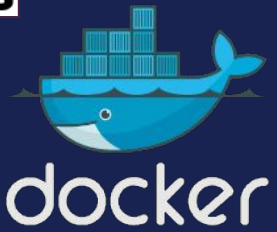
Debuggability & Control - Styx

- Execution control
- **Execution information**
  - Timeline
  - Google Cloud Logging



Debuggability & Control - Styx

- Execution control
- Execution information
- **Execution isolation**
  - Docker

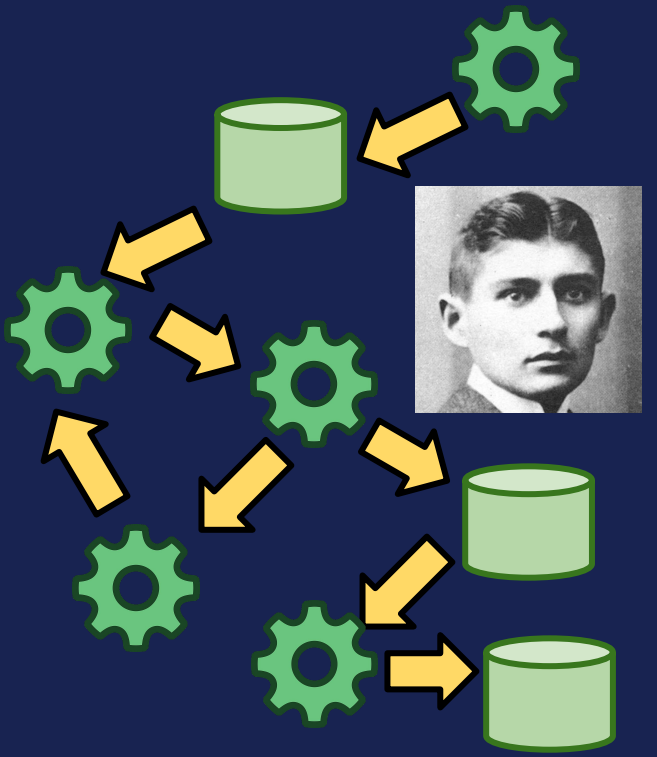


Challenges and the path to victory...

1. Early Warning — Datamon - Data monitoring
2. Debuggability & Control — Styx - Scheduling and control
3. Automate Capacity — GABO - Event Delivery

Automate Capacity - GABO/Event Delivery

- Complex and manual config



Automate Capacity - GABO/Event Delivery

- Complex and manual config
- Pubsub & Dataflow streaming

Automate Capacity - GABO/Event Delivery

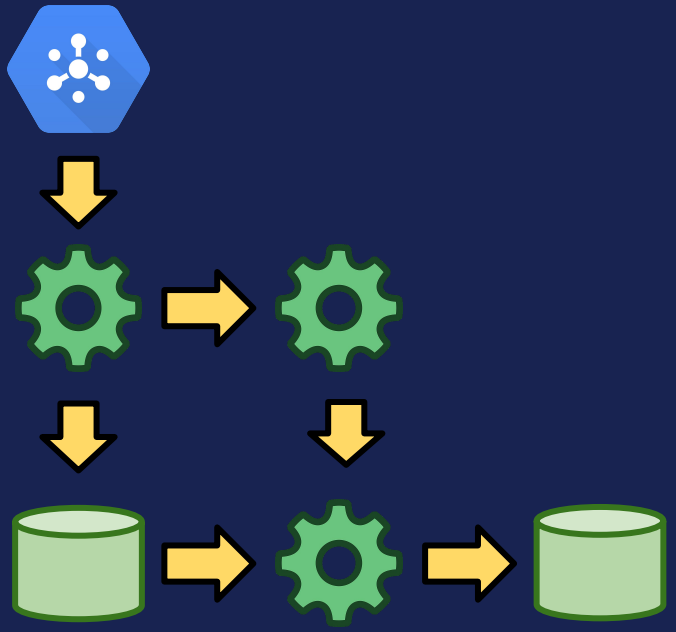
- Complex and manual config
- Pubsub & Dataflow streaming
- Pubsubs at scale

Automate Capacity - GABO/Event Delivery

- Complex and manual config
- Pubsub & Dataflow streaming
- Pubsubs at scale
- Dataflow streaming

Automate Capacity - GABO/Event Delivery

- Complex and manual config
- Pubsub & Dataflow streaming
- Pubsubs at scale
- Dataflow streaming :-(
- 2 micro services + 1 Map/Reduce job



Automate Capacity - GABO/Event Delivery

- Complex and manual config
- Pubsub & Dataflow streaming
- Pubsubs at scale
- Dataflow streaming :-(
- 2 micro services + 1 Map/Reduce job
- Autoscaling & The Stuffer

Summary

- **Make sure you have the right tools to deal with data incidents**
  - Make sure you have time to implement the tools you need

- **Remember that your capacity model can fail at larger scale**
  - Keep track of your scale and Automate, automate, automate...

Thank you!

matti@spotify.com

Want to join the band?
http://spoti.fi/jobs