Scaling Data Reliably: A journey in growing through data pain points

Miriah Peterson

Lead Engineer

SoyPete Tech

Do you have Reliable Data?

- Broken Dashboard?
- Missing Data Views?
- Airflow Job didn't run?
- Data missing in a table?
- Duplicate data in a table?
- API unavailable?
- Training job failed?

Data Downtime

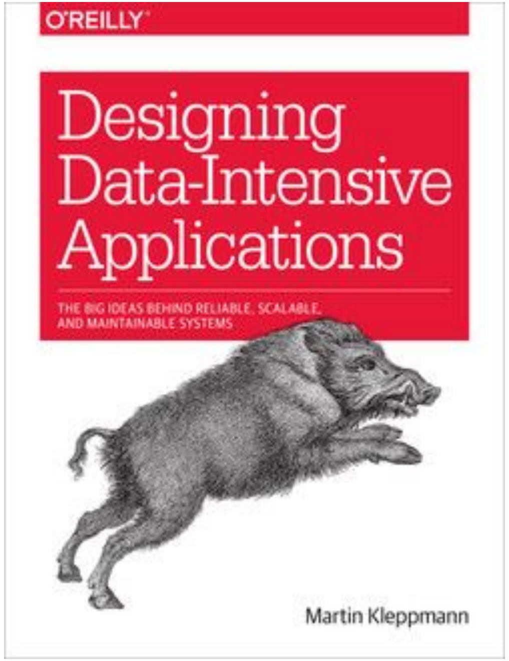
Data downtime refers to periods of time when your data is partial, erroneous, missing or otherwise inaccurate.

The Rise of Data Down Time, Barr Moses

What happens when data is down?

- Out of date dashboards
- Broken ML trainings
- Stopped financial operations

The Fundamentals


O'REILLY
Designing Data-Intensive Applications
THE BIG IDEAS BEHIND RELIABLE, SCALABLE, AND MAINTAINABLE SYSTEMS
Martin Kleppmann

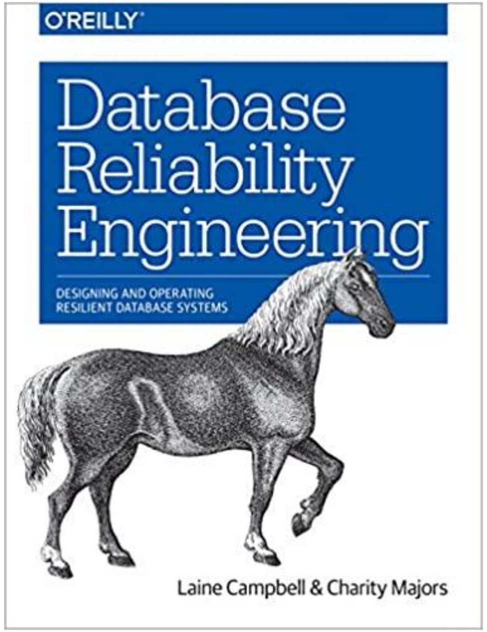
[Designing Data-Intensive Applications - Martin Kleppmann]

- Reliability
- Maintainability
- Operability

Reliable Data

Systems without failures, although robust, become **brittle and fragile**. When failures occur, it is more likely that the teams responding will be unprepared, and this could dramatically increase the impact of the incident.

- *Database Reliability Engineering, Laine Campbell & Charity Majors*


O'REILLY
Database Reliability Engineering
DESIGNING AND OPERATING RESILIENT DATABASE SYSTEMS
Laine Campbell & Charity Majors

Let’s talk about minimizing downtime

Lifecycle

Failure Points

Reliability Tools

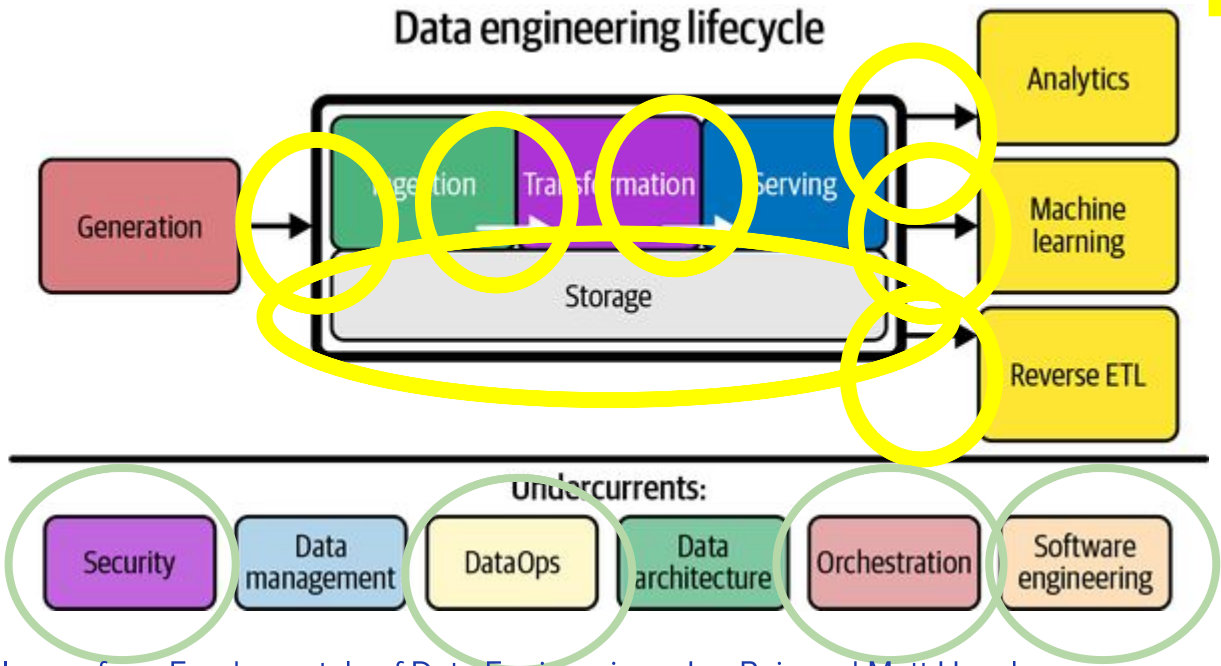


Image from Fundamentals of Data Engineering - Joe Reis and Matt Housley

Data Reliability Engineering

What is Data Reliability Engineering?

“I see Data Reliability Engineering as a **natural extension** of the data team. … Data Reliability Engineering means treating data quality like an **engineering problem**. It’s applying applications and tools to see that data stays for the variety of application use across the business.” — Egor Gryaznov, ¨Data Engineering Podcast,¨ episode 224[2]

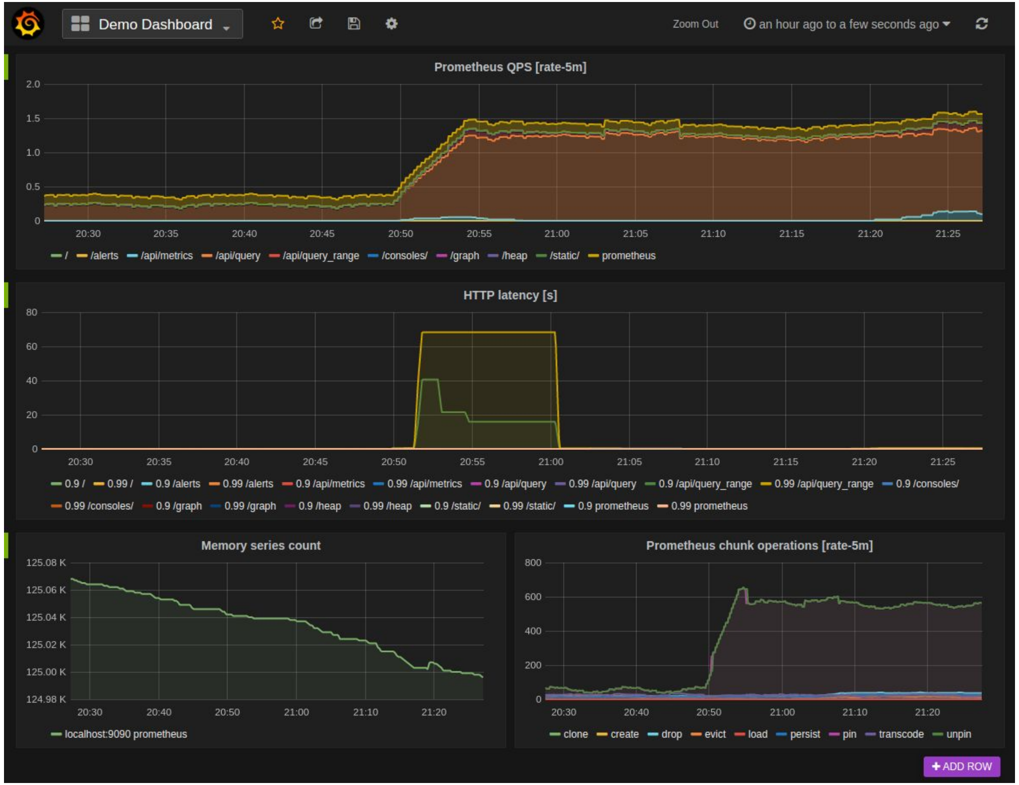
Measure our data

- Volume
  - How much data flows through your streams and apis?
  - How much data is added to the warehouse?
- Variety
  - Are all of the sources serving data?
- Veracity
  - Are the insights in line with the expected behavior?
- Value
  - Is all my data being used?
- Velocity
  - What is the throughput of my dataflows?

Data Service Metrics

- Latency
- Traffic
- Errors
- Saturation

Site Reliability Engineering - Betsy Beyer, Chris Jones, Jennifer Petoff and Niall Richard Murphy



Data Observability

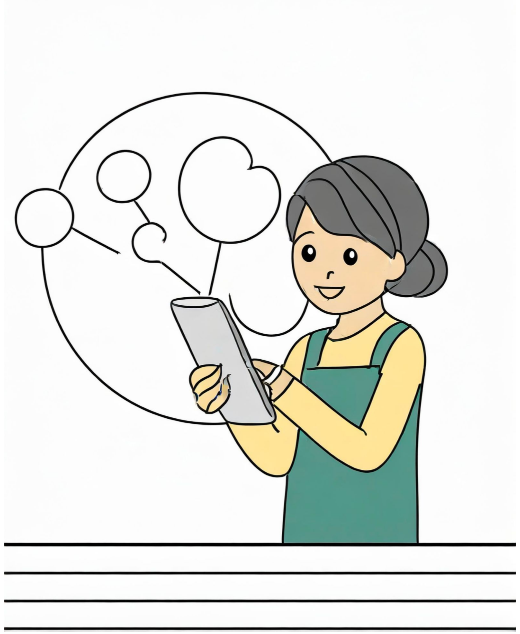
- SLAs
  - Dashboard has data with minimal freshness of 1 Day
- SLOs
  - Data Pipeline extracts data from source and completes transforms and analysis once a day
  - Can be manually triggered if there is an error
- SLIs
  - Errors in pipeline reported
  - Time out reported
  - Alerts of needed to be run manually

Data personas

Data Scientist

BI/Executive

ML application End User

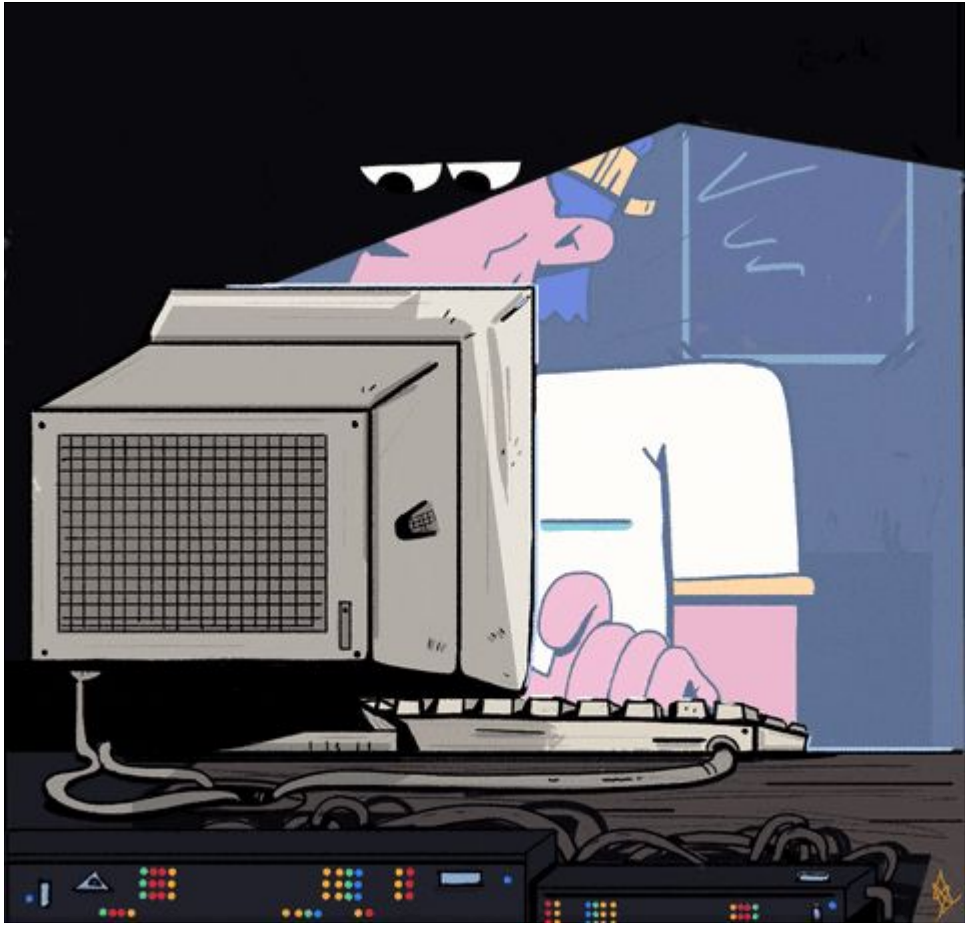


Images generated by adobe AI

Severity

Downtime != Outage

- Does it need to be addressed immediately?
- Can it wait until next business day?
- Can it wait until the next sprint?



Data Reliability Engineer

- Are you building tools to enable practitioners?
- Are you trying to automate your infra setup?
- Are you creating a platform for ML, AI, Analytics, etc?
- Are your regularly interfacing with a cloud infrastructure, operations, or SRE team?
- Are you struggling with downtime and you want to improve?

Data Reliability Engineer

The Data Reliability Engineer is the bridge between the Software Stack and the Data Stack

Images generated by adobe AI

Conclusion

- We all experience downtime
- Data Reliability Engineering is how we remedy it
- The larger our organization the more important metrics and quantization is for understanding our Reliability
- We need to create SLAs for the end users
- Use appropriate urgency when you remedy downtime

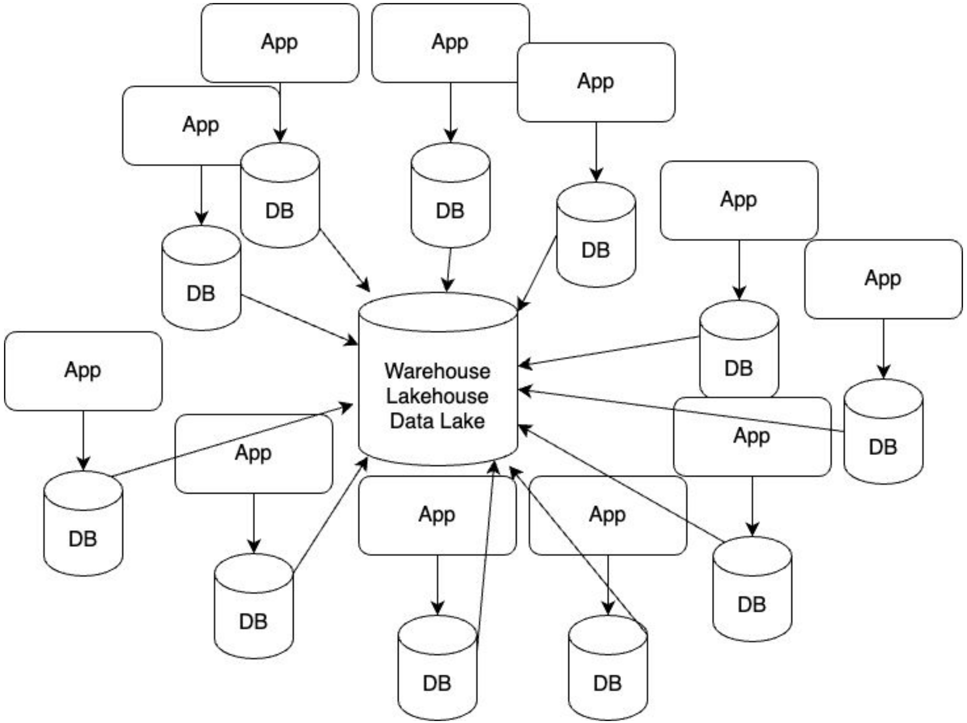
Thanks!

- Twitter: @captainnobody1
- Twitch: @soypete01
- LinkedIn: Miriah Peterson

References

- SRE books
- Data downtime
- Designing Data-Intensive Applications - Martin Kleppmann
- Database Reliability Engineering, Laine Campbell & Charity Majors
- Joe Reis and Matt Housley
- Data Reliability Podcast
- Intro to Data Reliability Engineering

What happens with microservices



Data System

Instead of designing a pipeline design a system

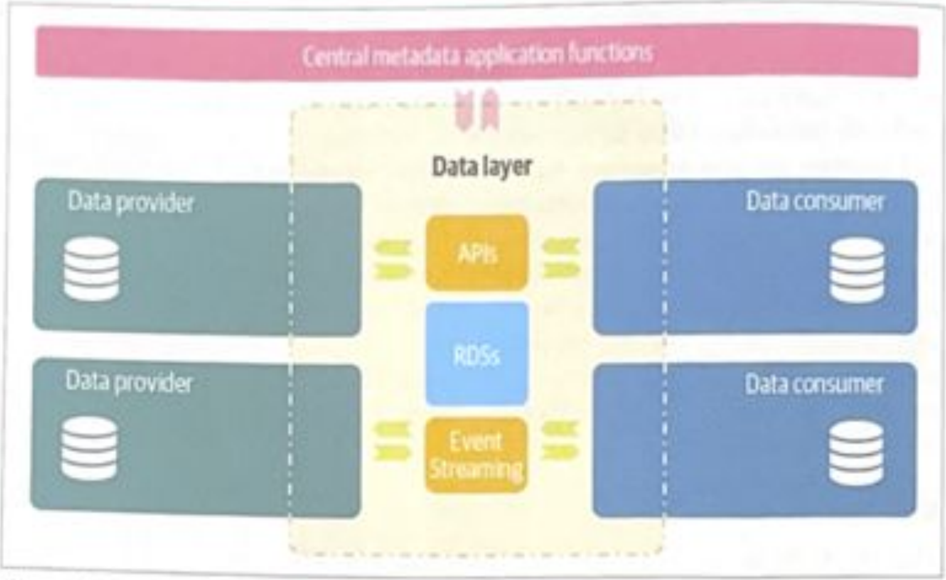


Figure 2-14. 1,000-foot view of the three different architectures and metadata.

https://www.oreilly.com/library/view/data-management-at/9781492054771/