Modern Data Team as a Team Sport

and how to be multi-sport ready with PipeRider

CL Kao, CEO & Co-founder

clkao@infuseai.io
About InfuseAI

Manage the complexity of AI workflow

Featured in “Market Guide for Artificial Intelligence Startups, GCR 2022”

Trusted by research institutes and leaders in sectors including FSI, manufacturing, and healthcare
Overview (expanded)

1. Structuring Data Teams
   a. Data being the nexus of High Agency Organization
   b. High Agency Individuals to High Agency Teams
   c. What we learned from org structure & tools in software engineering

2. Data Observability & Reliability
   a. The emerging data stack and self-service
   b. How things can break and hurt Trust
   c. Observing the many moving parts

3. PipeRider: Open Source Data Reliability Toolkit
   a. Warehouse-native profiling & testing
   b. Capturing anomalies and unintended data impacts
   c. PipeRider Cloud & Design Partnership
Overview

1. Structuring Data Teams
2. Data Observability & Reliability
3. PipeRider: Open Source Data Reliability Toolkit

https://github.com/infuseai/piperider
About CL Kao

- Open Source developer since 1997
- Creator of pre-Git distributed version control systems, used by Apple
- Started the g0v (gov-zero) civic tech movement in 2012
- Started InfuseAI in 2018
- Father of two
I think about how people collaborate a lot...

The gov-zero (g0v) community after a decade

The g0v Manifesto

We come from everywhere

We are a polycentric community of self-organized contributors

We are citizens collaborating to bring about change

We live open-source

We have fun and want to change the status quo

We are you

https://g0v.tw/intl/zh-TW/manifesto/en/
Photo by: Daisuke Chen
High Agency Movement - Through (Open) Data

Campaign Finance Filing

- Turning Scanned PDF into captcha game
- Crowd-source digitalization by 10k people
- Meshed with Company registry to find conglomerate cluster
- Analysis of money in politics

- After 5-years, law amended to mandate disclosure of campaign finance digitally
Structuring Data Teams
High Agency Individuals

“Relentlessly resourceful”

Cultivate for Collective High Agency, through:

- Culture
- Supporting Infrastructure
High Agency Reinforces Itself

When:

- Intermediate Outcome Achievable by Individuals
  - Data in this case
- Bound by Shared Goals
- Frictionless technical layers for sharing Code, Data, Ideas
How Software Tooling Enabled Modern Development

- “svk” - predates git that enables decentralized collaboration of software development and version control
- Adopted by Apple to manage WebKit forks for iPhone
- “forking” by default, which is permissionless, nowadays commonly knowns as “pull requests”
Feature Teams vs Empowered Product Teams

Build Features          Solve Problems
Design & Code          Discover Best Solutions
Output                 Outcome

EMPOWERED
ORDINARY PEOPLE,
EXTRAORDINARY PRODUCTS
Data Team Organization Spectrum

- Centralized
  - CoE / Consultant
- Embedded
- Polycentric
  - Decentralized Analyst + Data Platform Team
  - “Product Data Science”
  - “Data Business Partners”

Bottom Line: Data Roles shouldn’t be just taking requests to build dashboards

https://www.getdbt.com/data-teams/centralized-vs-decentralized/

https://djpardis.medium.com/models-for-integrating-data-science-teams-within-organizations-7c5afa032ebd
Data Observability & Reliability
Modern Data Stack & Self-Service
Modern Data Stack & Self-Service

Paradigm: ETL to ELT (in-warehouse transformations)

New Challenges

- Metadata of 10x more tables
- Data governance issues
- Best practices / Conventions
- Always on the way to data-la-la-land:
  
  Everything is perfectly and flexibly Defined
And this leads us to...

“Propensity_to_purchase depends on...”
Data Observability

Continuously monitoring data for anomalies:

- Distribution Change
- Duplications
- Stalled Data
- Missing or Invalid Schema
Data Reliability

Mitigating the issues

- Rules, Constraints
- Alerting
- Out of Bound Process of Bad Data

Part of the development cycle, anticipating changes from code & data
How Big is the Problem?

- 2017: MIT Sloan Review: cost of bad data to be 15% to 25% of revenue for most company
- 2022: Unity lost $110m due to bad data ingestion, disclosed in earnings call

The first was a fault in our platform that resulted in reduced accuracy for our Audience Pinpointer tool, a revenue expensive issue given that our Pinpointer tool experienced significant growth post the IDFA changes.

The second is that we lost the value of a portion of our data, training data due in part to us ingesting bad data from a large customer. We estimate the impact to our business of approximately $110 million in 2022 with no carryover impact to 2023.

– Unity Software Inc. (NASDAQ:U) Q1 2022 Earnings Call Transcript
PipeRider: Open Source Data Reliability Toolkit
Piperider

Data Reliability Toolkit

👋 Open Source
🍰 Minimum Setup
🛸 Non-invasive
Warehouse-Native Data Profiling

Install PipeRider

```bash
~ $ pip install -U piperider
```

Analyze a datastore in seconds!

```bash
~ $ cd dataproject
# initialize PipeRider
dataproject $ piperider init
# profile and test your data
dataproject $ piperider run
```
# PipeRider: Data Report

## Data Composition

<table>
<thead>
<tr>
<th>Column</th>
<th>Valid (%)</th>
<th>Invalid (%)</th>
<th>Missing (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>100.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Valid</td>
<td>100.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Invalid</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Missing</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

## Numeric Composition

- **Zeros**: 20.9% (40)
- **Positives**: 79.1% (151)

## Numeric Statistics
Comparing Profile Reports

1 Profile Run:
- shows you statistics
- suggests assertions to enforce

2 Profiles: difference, what changed

3+ Profiles: trends, anomalies
Catch Data Impacts during Continuous Integration

Comparison result help to know the impact for the PR.
As your Data Co-Pilot, Reliability from Day 1
PipeRider Cloud (Sign up for early access: piperider.io)

- Report Sharing
- Anomalies & Alerts
- Seamless CI integrations
- Automated Downstream Modeling
Conclusion
Conclusion: Optimize for Collectively High Agency

Data as a Team Sport, and be Multi-Sport Ready

Data-Powered Social Movements, Empowered Product Teams Show us:

- (Tech/Culture) Infrastructure + High Agency
Conclusion: Optimize for Collectively High Agency

Structuring Data Teams

Deep in Product/Business outcome, Aligned in practices

Data Observability & Reliability

Invest in data infrastructure promoting self-service, assist collaborative practices

Moving Fast with Confidence

PipeRider - https://github.com/infuseai/piperider
Q&A

We’d love to chat! How do you organize your data teams?

hi@infuseai.io

Star us: https://github.com/infuseai/piperider