InfuseAI

Building Trustworthy Pipelines for AI and Analytics

Chia-Liang Kao
Co-Founder & CEO
Hi, I am CL Kao (@clkao)

• Open Source developer since 1997
• Created one of the distributed version control systems before Git
• Started the g0v(gov-zero) civic tech movement in 2012
• Started InfuseAI in 2018
• Father of two
About InfuseAI

trusted by research institutes and leaders in sectors including FSI, manufacturing, and healthcare

100+ projects our product manage for clients

500+ GPUs our product manage for clients

500 Invested by 500 Startups

1st Private Enterprise Tech Innovator KPMG, 2021

AI Infrastructure Alliance
Making use of Data is not as simple
It usually involves Pipelines
.. Or Complicated Pipelines
Things can go wrong easily
Distribution Change
Missing or Invalid Values
Pipeline-Wide Observability
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
Pervasive Label Errors in ML Datasets Destabilize Benchmarks (2021)
- Curtis G. Northcutt, Anish Athalye, Jonas Mueller

“We estimate an average of 3.4% errors across the 10 datasets, where for example 2,916 label errors comprise 6% of the CIFAR-100 test set and ~390,000 label errors comprise ~4% of the Amazon Reviews dataset”

![Images of MNIST, CIFAR-10, CIFAR-100, Caltech-256, ImageNet, QuickDraw with labels and corrections](https://labelerrors.com)
Data-Centric AI requires Iterating Data
Common Data Quality Issues

- Distribution Changes
- Stalled Data
- Missing or Invalid
- Schema or Semantic Changes
Data \{Quality, Documentation, Cleaning\} Tools

- dbt
- Pandas Profiling
- Great Expectations
- TDDA
- dedupe
- cleanlab
- dataprep
dbt vs Great Expectations

**dbt**
- Great if already modeling with dbt
- Tests are SQL-based and run inside warehouse
- Lots of built-in tests and extensions

**Great Expectations**
- Python-based
- Unified api for different backend (in-memory, spark, SQL)
- Flexible with how to orchestrate
Data Version Control Tools

- git-lfs
- dvc
- ArtiVC.io
- (managing files by yourself)
Is Testing Your Data Like Code Enough?
Data Quality Tools on Pipeline
There are More than One Way to Do it

github.com/spbail/dag-stack
dagster

- dbt: Test integrity of transformations e.g. no fan-out joins, no NULL columns, etc.
- Use off-the-shelf methods for complex tests, e.g. distributions of values - and generate Data Docs
PipeRider: Pipeline-Wide Data Quality

💰 PipeRider

[Diagram showing data flow and money bag]
PipeRider

✈️ Open Source
🍰 Minimum Setup
🛸 Non-invasive
Example: Real-time Antigen Inventory

- Downstream:
  - Analysis for antigen demand by region
  - Real-time dashboards
  - Prediction of inventory
Initialize PipeRider project

$ piperider init --from-dbt

Configured data source: snowflake (via dbt project nhi_fst)

1 Source added.
2 Models added.

Config written: PipeRider.yml

$ piperider run --generate-report

Report available at output/nhi-project-2022-05-17.html
Example: Pull-Request for Modeling Changes
Example: PR Impact with Lineage

- ** nhi_airbyte_raw Nhi_fst**
  - freshness
  - all columns pass

- **inventory_event**
  - all columns pass

- **facility**
  - some columns have unseen missing values
  - schema changes detected
  - see details

- **Inventory Prediction Model**
  - not up-to-update
  - impact warning
# Example: Data Profile Changes

**Table:** Facility

<table>
<thead>
<tr>
<th>Column</th>
<th>Type</th>
<th>Distribution</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>code</td>
<td>text</td>
<td>no missing values</td>
<td></td>
</tr>
<tr>
<td>name</td>
<td>text</td>
<td>no missing values</td>
<td></td>
</tr>
<tr>
<td>New admin1</td>
<td>text</td>
<td>4965 missing values</td>
<td>![Add as test]</td>
</tr>
<tr>
<td>New admin2</td>
<td>text</td>
<td>10 missing values</td>
<td>![Review test]</td>
</tr>
<tr>
<td>address</td>
<td>text</td>
<td>no missing values</td>
<td>![Add as test]</td>
</tr>
<tr>
<td>longitude</td>
<td>double precision</td>
<td>no distribution change (show)</td>
<td>![Review test]</td>
</tr>
<tr>
<td>latitude</td>
<td>double precision</td>
<td>no distribution change (show)</td>
<td>![Review test]</td>
</tr>
<tr>
<td>phone</td>
<td>text</td>
<td>no missing values</td>
<td>![Add as test]</td>
</tr>
<tr>
<td>comment</td>
<td>text</td>
<td>no missing values</td>
<td>![no distribution change (show)]</td>
</tr>
</tbody>
</table>

- **Test Coverage Change (80% → 65%)**
- **Record Count Change (3%)**
- **Schema Change: 2 new columns**

- **Distribution change > 5% threshold**
  - longitude
  - latitude

- **Review test**
Modern Monitoring and Diagnosis for Pipelines

Be the first to know the impact of every change in your pipelines and avoid costly data incidents.

https://PipeRider.io
Open Source Standard for Managing the Complexity of AI Workflows

Our Goal

Our Products

PipeRider
pipeline-wide change management

PrimeHub
manage ML lifecycle

Crane
manage container images

ArtiVC
large file data version control
Conclusion

⚡ Trust => Efficiency

孔雀 Data Stack is Polyglot

🐒 Pragmatic Observability Across Stack
Q&A
Manage the Complexity of AI Workflows Seamlessly

hi@infuseai.io

Thank to lovely icons makers:
- Operating model icons created by inipagistudio
- Pipe icons created by Flat Icons
- Folder icons created by Freepik
- Dashboard icons created by juicy_fish
- Chemistry icons created by Freepik