



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

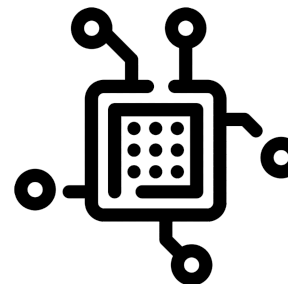
About InfuseAI



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



AI Infrastructure Alliance



Invested by
500 Startups

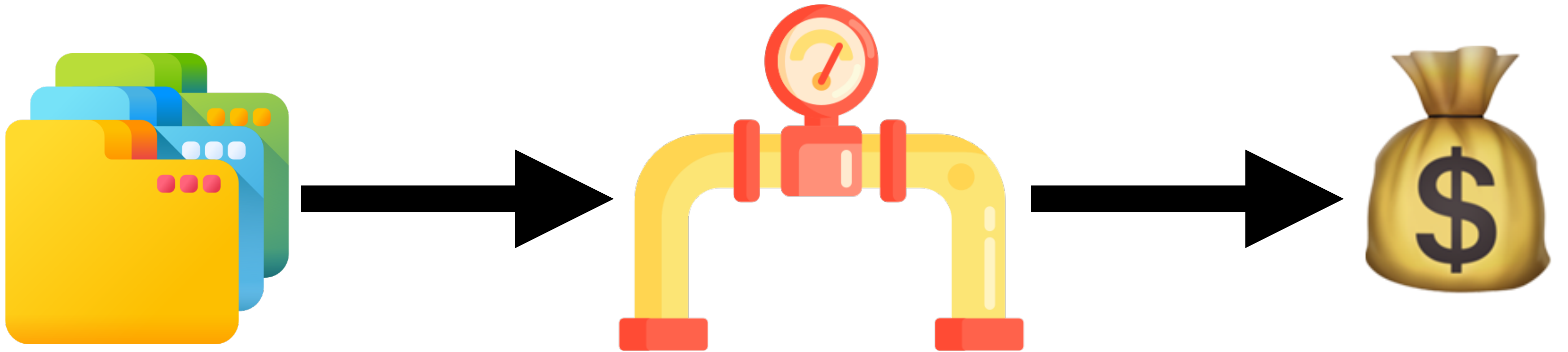


Private Enterprise
Tech Innovator
KPMG, 2021

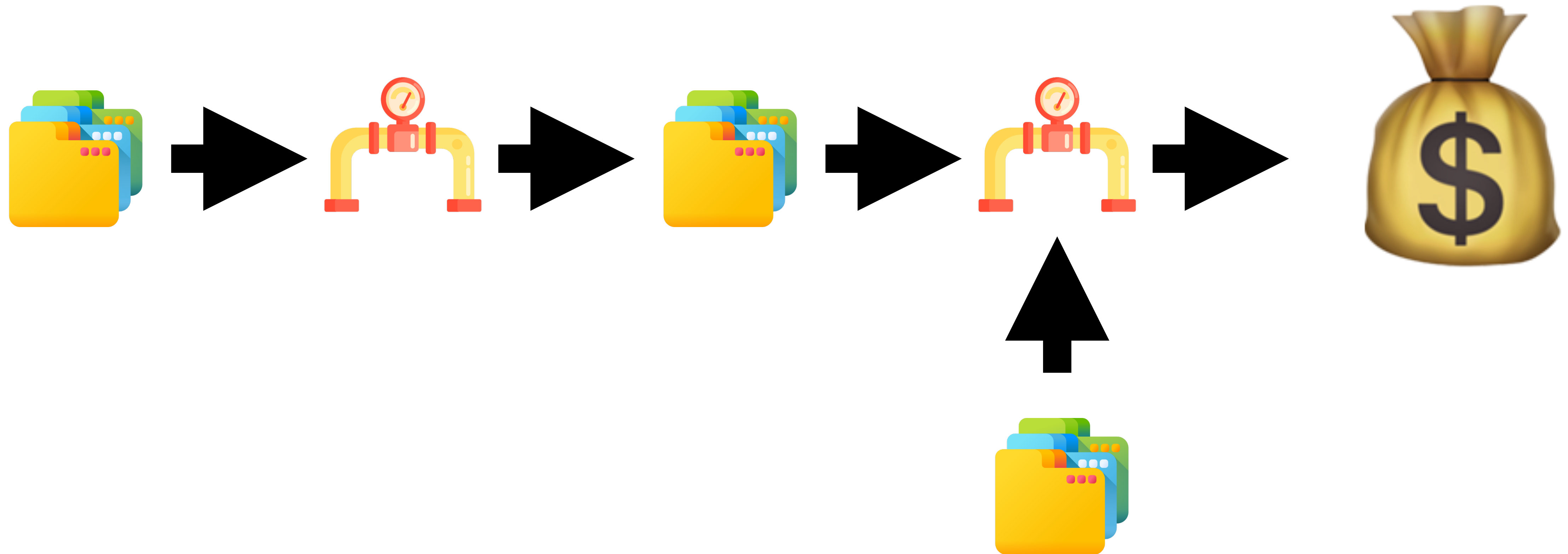
Making use of Data is not as simple



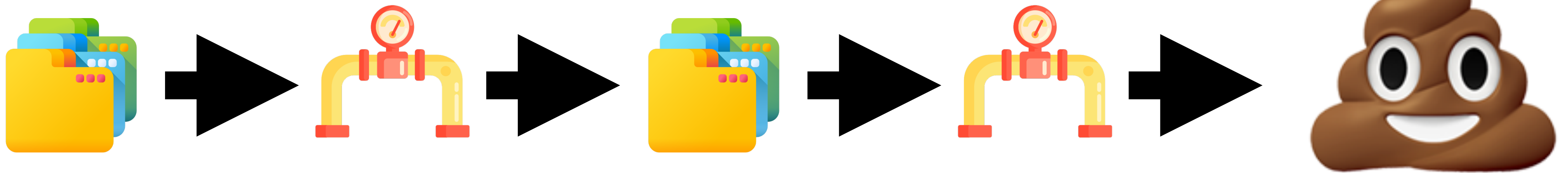
It usually involves Pipelines



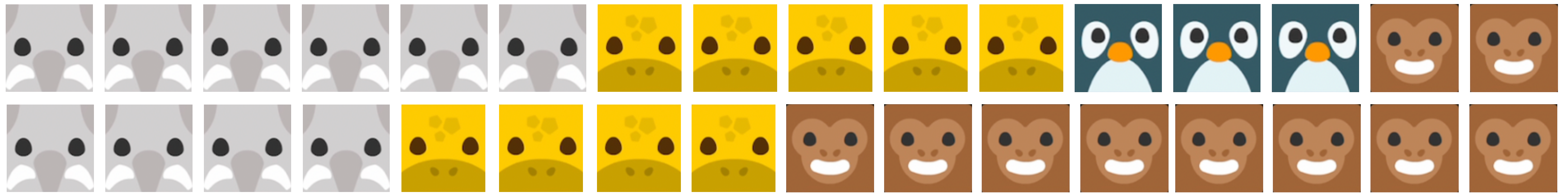
.. Or Complicated Pipelines



Things can go wrong easily



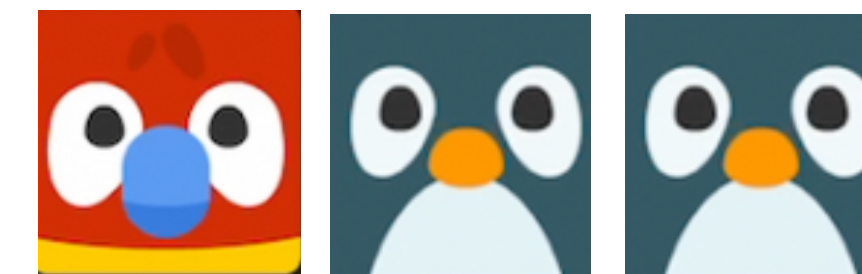
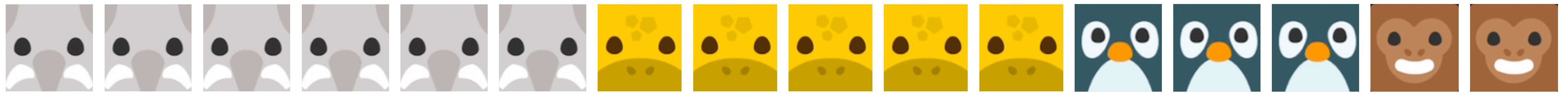
Distribution Change



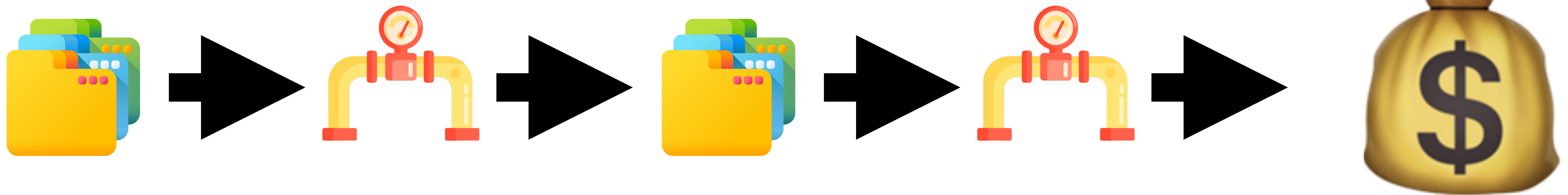
Incorrect Labels



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

How Big is the Problem?

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”

MNIST



given: 5
corrected: 3

CIFAR-10



given: cat
corrected: frog

CIFAR-100



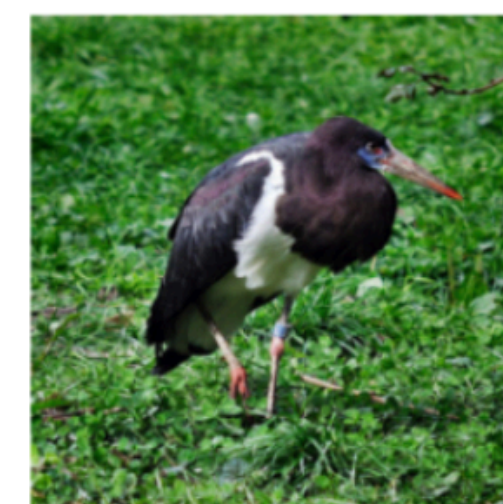
given: lobster
corrected: crab

Caltech-256



given: ewer
corrected: teapot

ImageNet



given: white stork
corrected: black stork

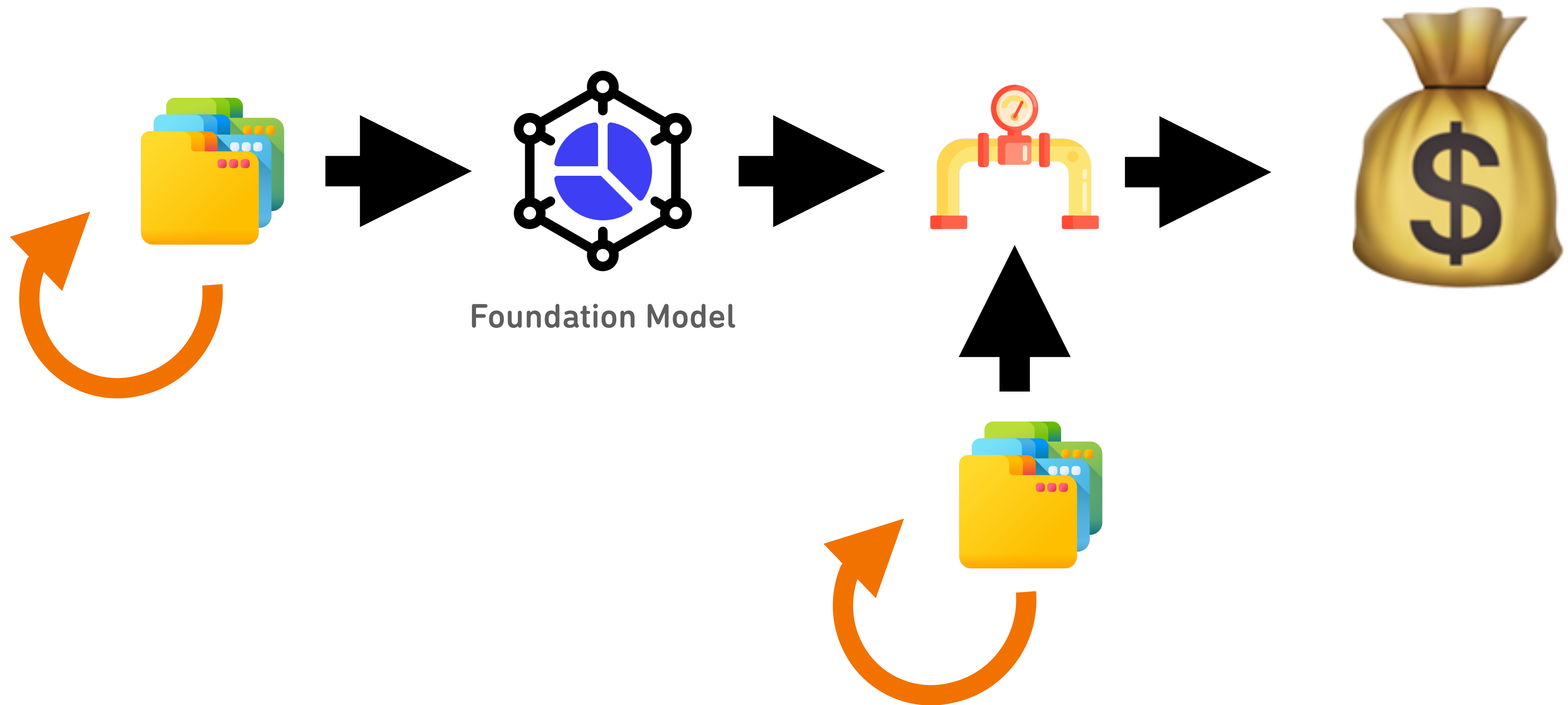
QuickDraw



given: tiger
corrected: eye

<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

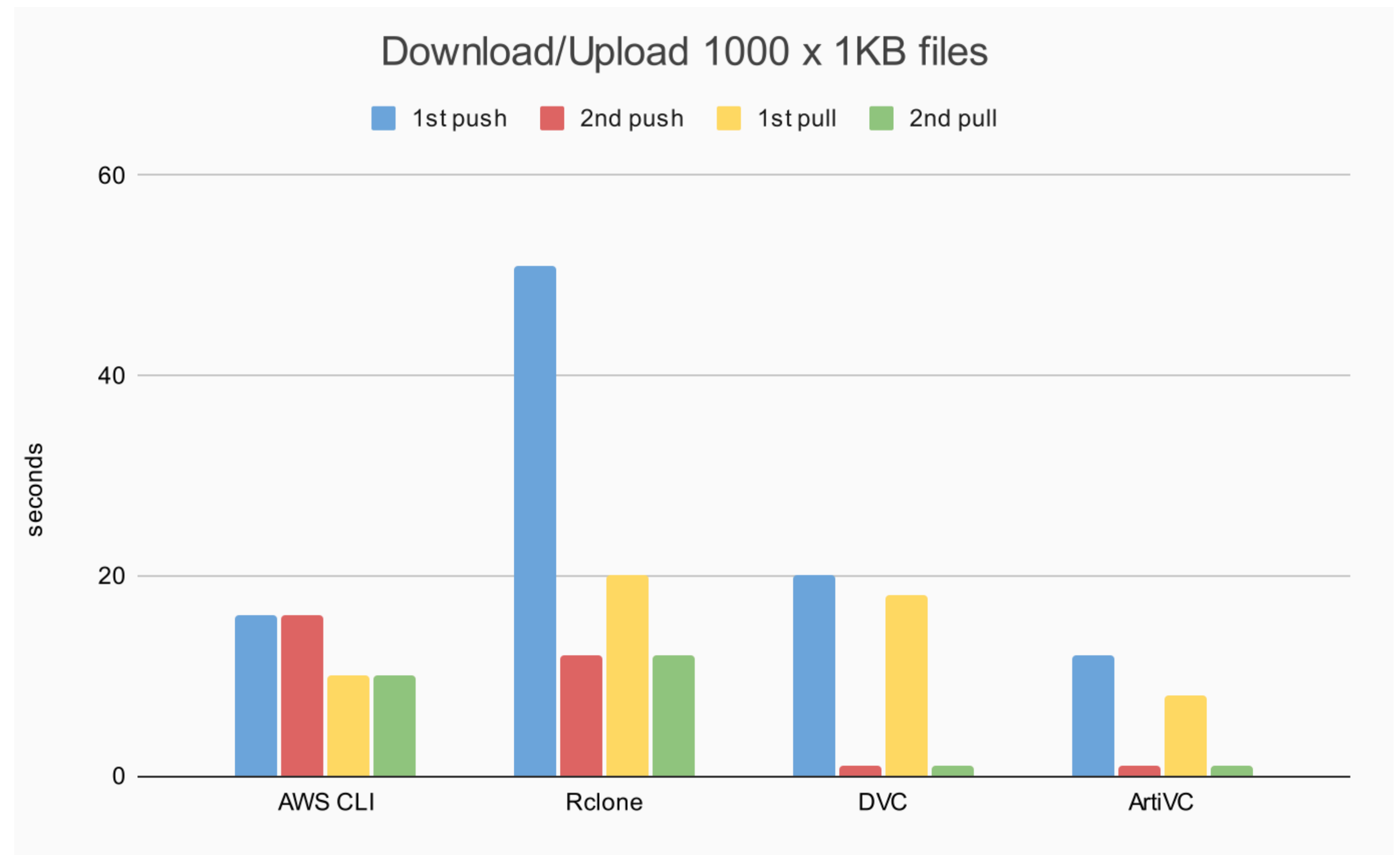
- Great if already modeling with dbt
- Tests are SQL-based and run inside warehouse
- Lots of built-in tests and extensions

Great Expectations

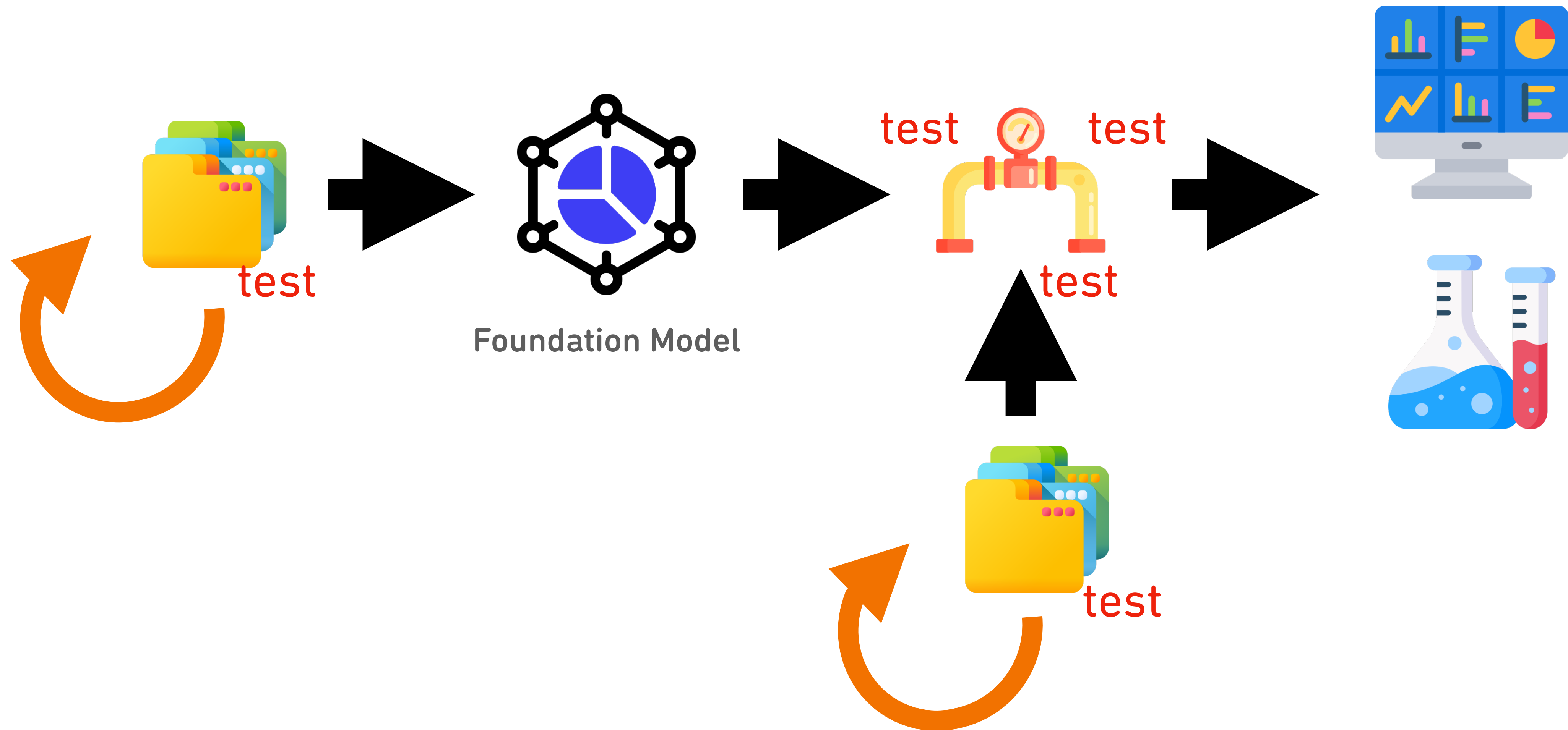
- Python-based
- Unified api for different backend (in-memory, spark, SQL)
- Flexible with how to orchestrate

Data Version Control Tools

- git-lfs
- dvc
- ArtiVC.io
- (managing files by yourself)



Is Testing Your Data Like Code Enough?

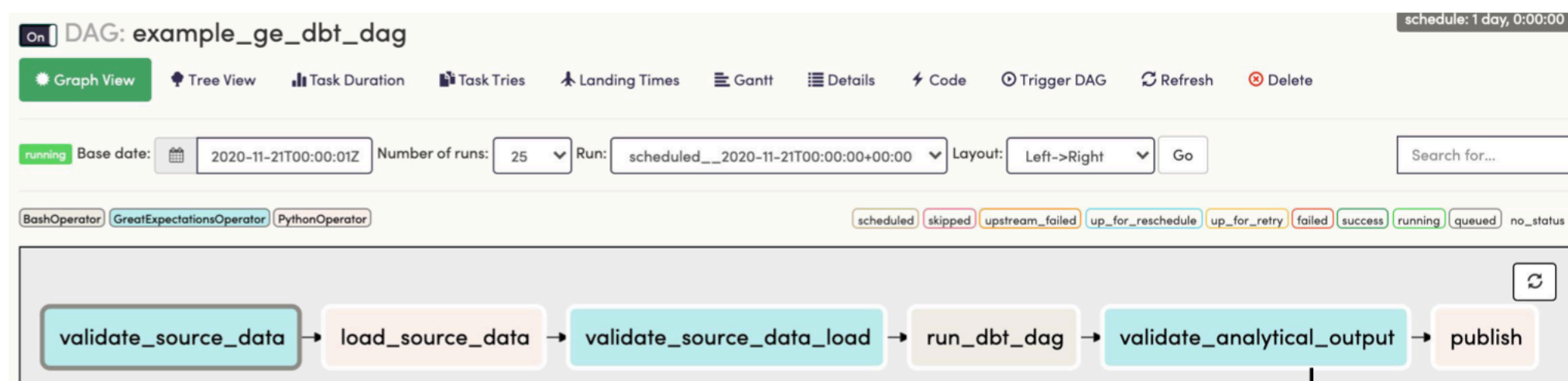


Data Quality Tools on Pipeline

There are More than One Way to Do it



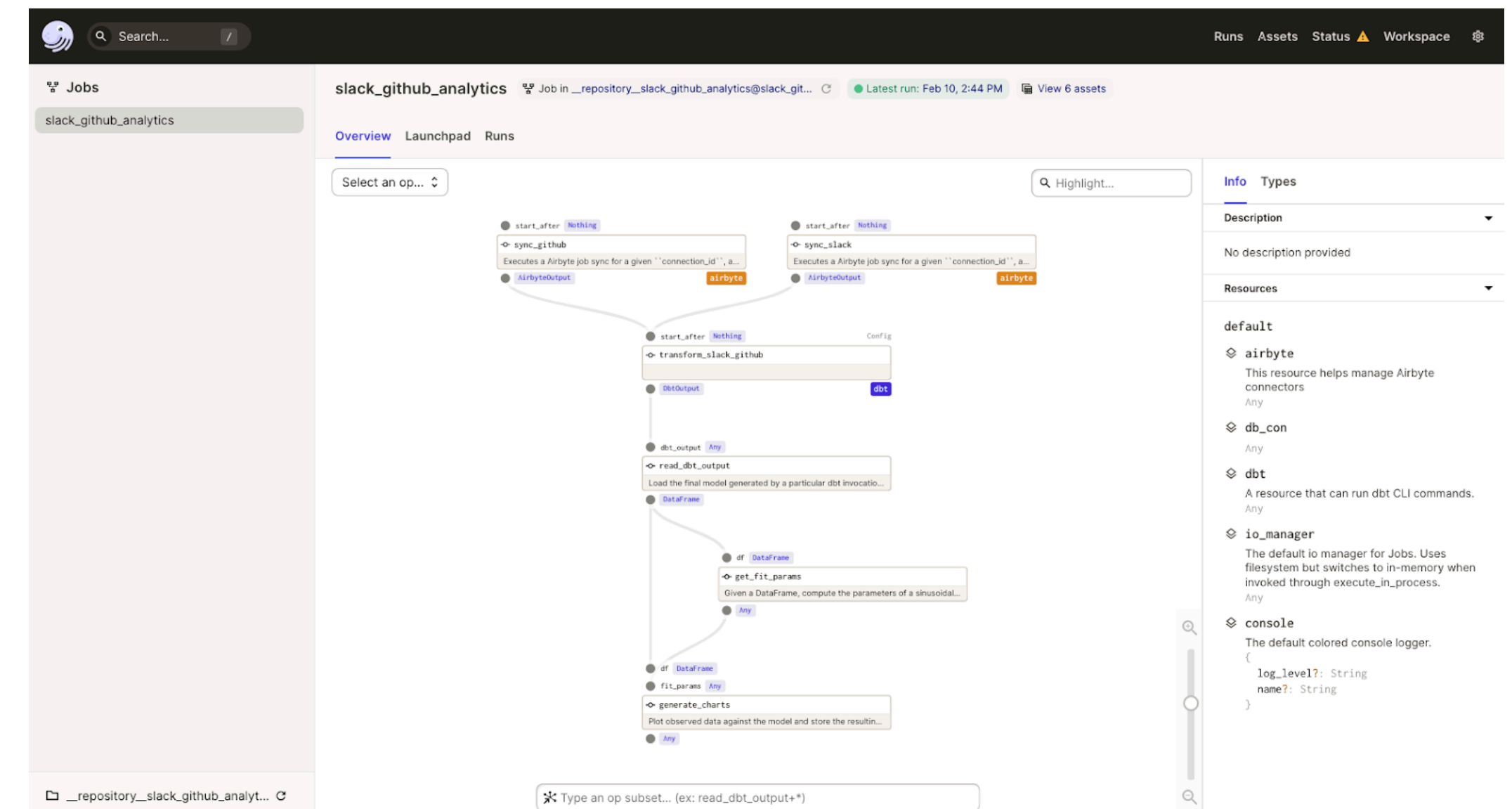
github.com/spbail/dag-stack



dbt Test integrity of transformations e.g. no fan-out joins, no NULL columns, etc.

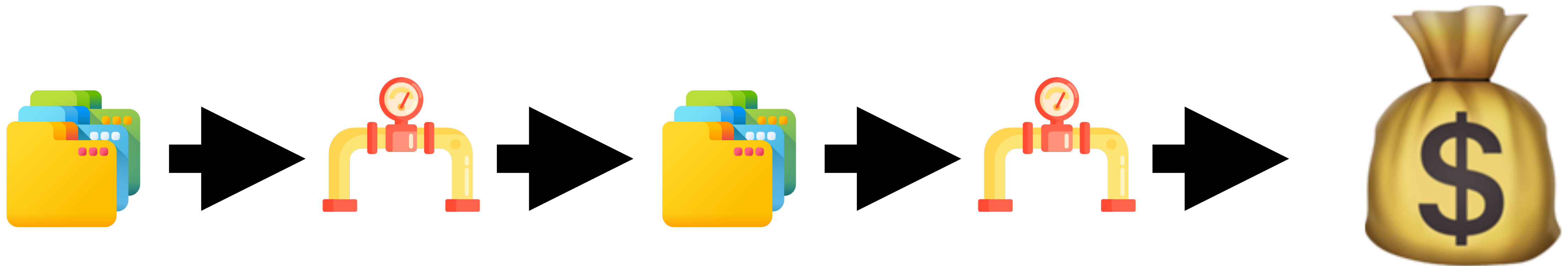
Use off-the-shelf methods for complex tests, e.g. distributions of values - and generate Data Docs

dagster



PipeRider: Pipeline-Wide Data Quality

 PipeRider





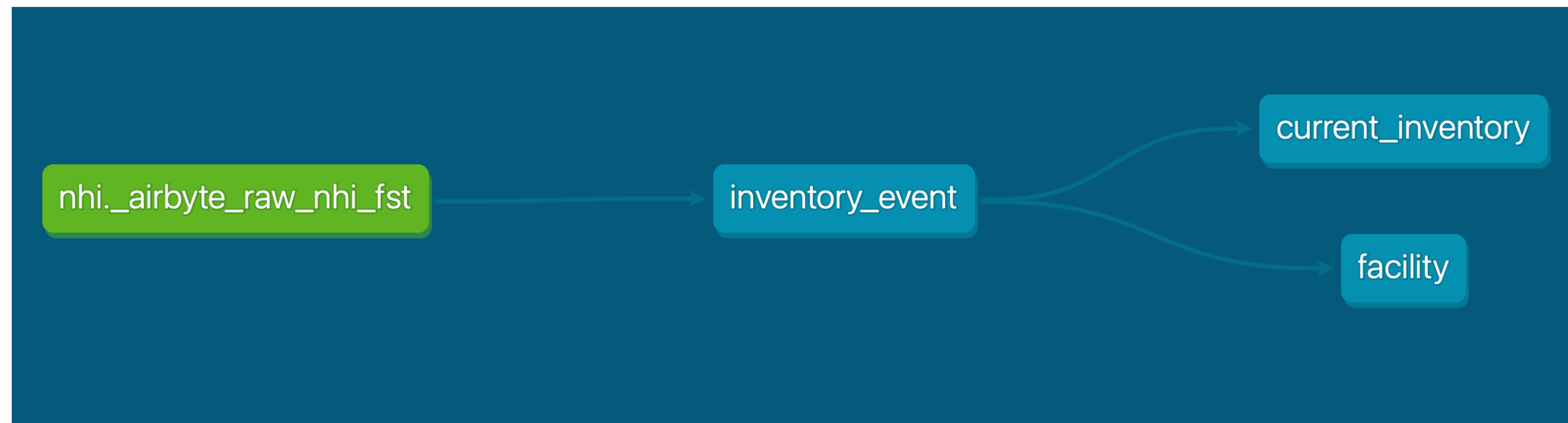
PipeRider

👐 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.
```

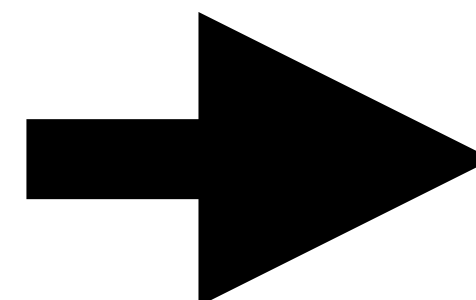
```
Config written: PipeRider.yml
```

```
→ $ piperider run --generate-report
```

```
Report available at output/nhi-project-2022-05-17.html
```

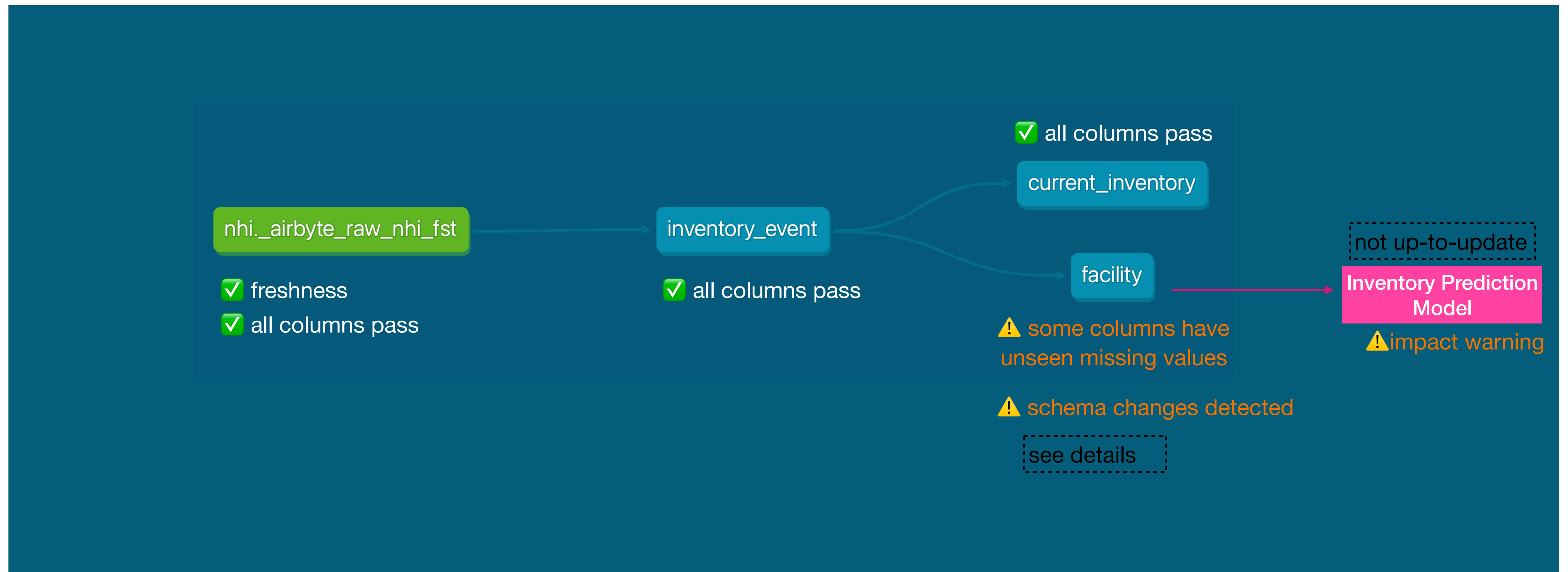

Example: Pull-Request for Modeling Changes

```
add admin1 & admin2 columns #3
Merged clkao merged 2 commits into main from admin 15 days ago
Conversation 0 Commits 2 Checks 1 Files changed 2 +10 -0
Changes from all commits File filter Conversations 0/2 files viewed Review changes
Filter changed files
models/nhi_fst
  facility.sql
  schema.yml
  models/nhi_fst/schema.yml
@@ -21,6 +21,14 @@ models:
21     tests:
22     - not_null
23
24 +   - name: facility
25 +     description: "All participating
26 +       facilities"
27 +     columns:
28 +       - name: admin1
29 +         description: "City / County"
30 +     tests:
31 +       - not_null
24 - name: current_inventory
25   columns:
26   - name: quantity
32 - name: current_inventory
33   columns:
34   - name: quantity
```



```
Test dbt models 46s
1 Run echo "::set-output name=test_result::$(dbt test)\n"
13 16:46:38 Found 3 models, 4 tests, 0 snapshots, 0 analyses, 380 macros, 0 operations, 0 seed files, 1 source, 0
    exposures, 0 metrics
14 16:46:38
15 16:46:39 Concurrency: 4 threads (target='dev')
16 16:46:39
17 16:46:39 1 of 4 START test dbt_utils_accepted_range_current_inventory_quantity__1000__0 . [RUN]
18 16:46:39 2 of 4 START test not_null_current_inventory_quantity ..... [RUN]
19 16:46:39 3 of 4 START test not_null_facility_admin1 ..... [RUN]
20 16:46:39 4 of 4 START test not_null_inventory_event_source_updated_at ..... [RUN]
21 16:46:39 3 of 4 FAIL 4965 not_null_facility_admin1 ..... [FAIL 4965 in 0.21s]
22 16:46:39 4 of 4 PASS not_null_inventory_event_source_updated_at ..... [PASS in 0.22s]
23 16:47:22 2 of 4 PASS not_null_current_inventory_quantity ..... [PASS in 43.20s]
24 16:47:22 1 of 4 PASS dbt_utils_accepted_range_current_inventory_quantity__1000__0 ..... [PASS in 43.20s]
25 16:47:22
26 16:47:22 Finished running 4 tests in 43.67s.
27 16:47:22
28 16:47:22 Completed with 1 error and 0 warnings:
29 16:47:22
30 16:47:22 Failure in test not_null_facility_admin1 (models/nhi_fst/schema.yml)
31 16:47:22 Got 4965 results, configured to fail if != 0
```

Example: PR Impact with Lineage



Example: Data Profile Changes

Table: Facility

⚠ Test Coverage Change (80% → 65%)

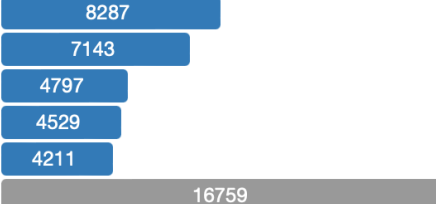
Record Count Change (3%)

Review test

Columns

⚠ Schema Change: 2 new columns

Add as test

COLUMN	TYPE		DISTRIBUTION
code	text	no missing values	no distribution change (show)
name	text	no missing values	no distribution change (show)
New admin1	text	⚠ 4965 missing values	
New admin2	text	⚠ 10 missing values	Add as test
address	text	no missing values	no distribution change (show)
longitude	double precision	no missing values	✖ distribution change > 5% threshold Review test
latitude	double precision	no missing values	✖ distribution change > 5% threshold Review test
phone	text	no missing values	no distribution change (show)
comment	text		

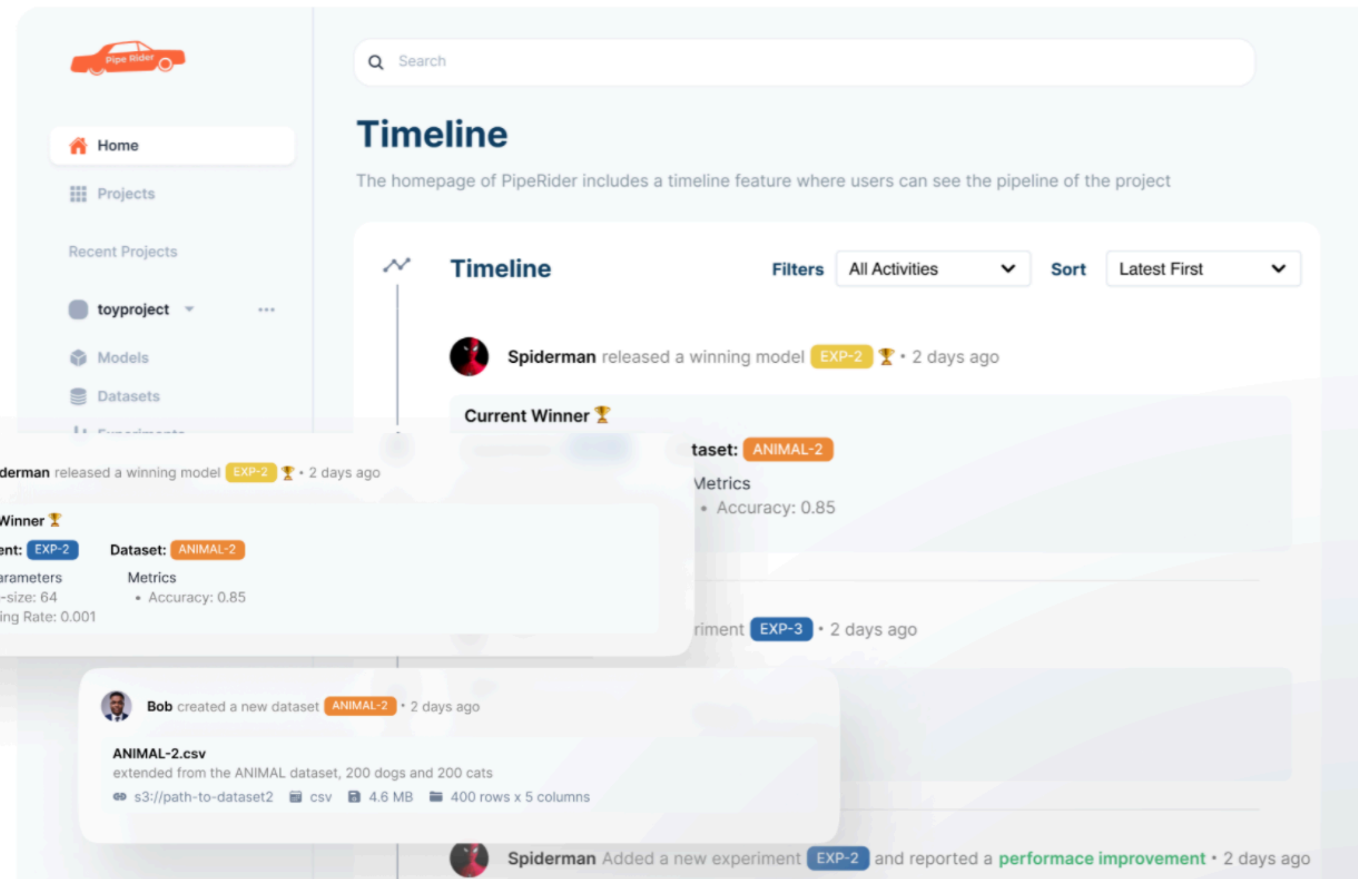
<https://PipeRider.io>



Request Early Access

Modern Monitoring and Diagnosis for Pipelines

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



Our Goal

Open Source Standard for Managing the Complexity of AI Workflows

Our Products



pipeline-wide change management



Conclusion

⚡ Trust => Efficiency

🦜 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](#)