Data Lineage

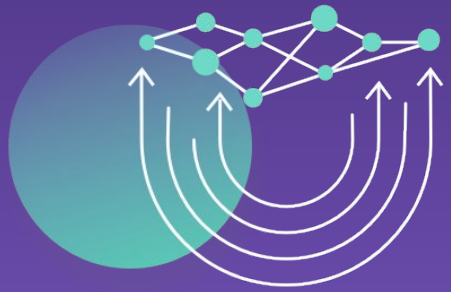
What is it, and why do we care?

Ross Turk

ASTRONOMER

ASTRONOMER

Working to advance
Apache Airflow



datakin

The principal drivers of
OpenLineage

The 2022 Airflow User Survey

Open until Friday, June 3

https://bit.ly/AirflowSurvey22

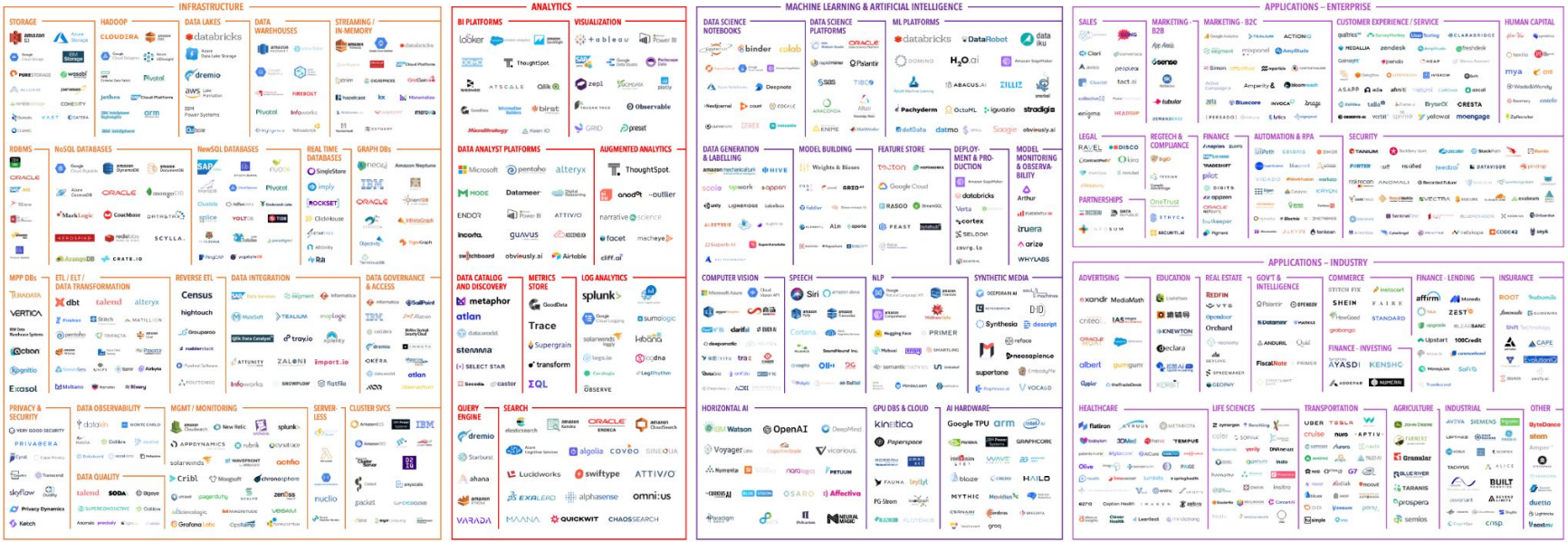
These are individual opinions, I don't speak for anyone else.

I am pretty sure my employer agrees with them (but I didn't ask...)

What creates the need for data lineage?

Context | What | Why | How

MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, AND DATA (MAD) LANDSCAPE 2021



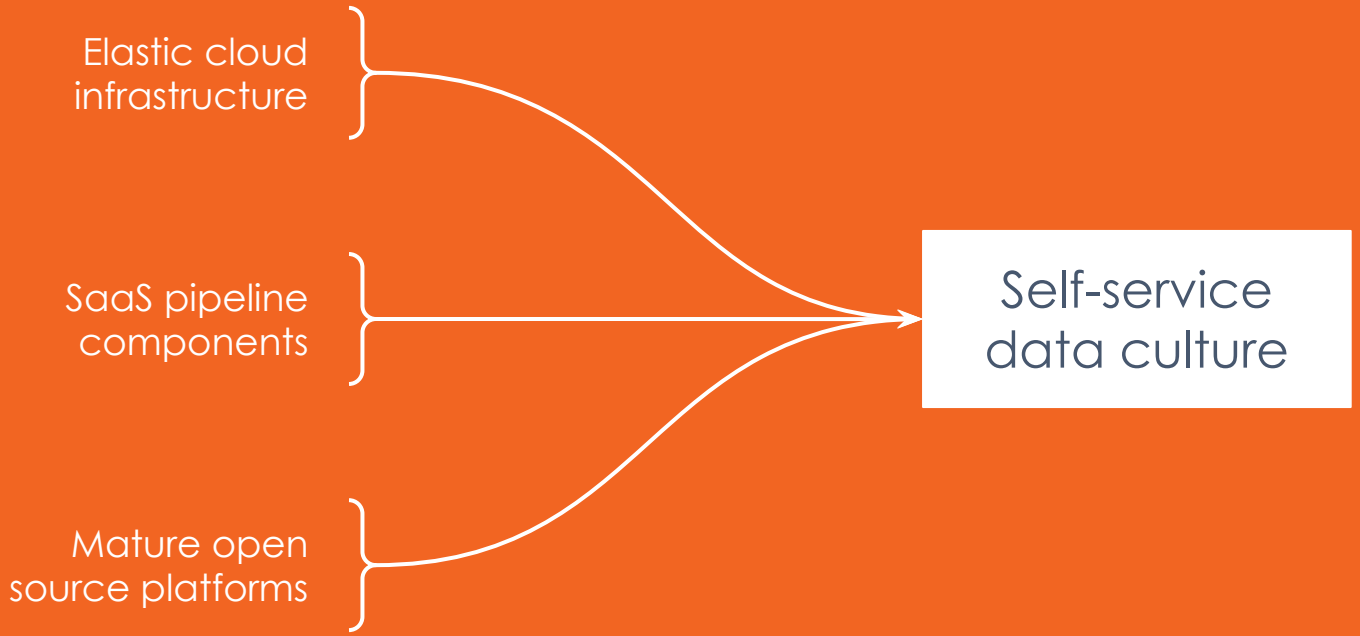
Version 3.0 - November 2021

© Matt Turck (@mattturck), John Wu (@john_d_wu) & FirstMark (@firstmarkcap)

mattturck.com/data2021

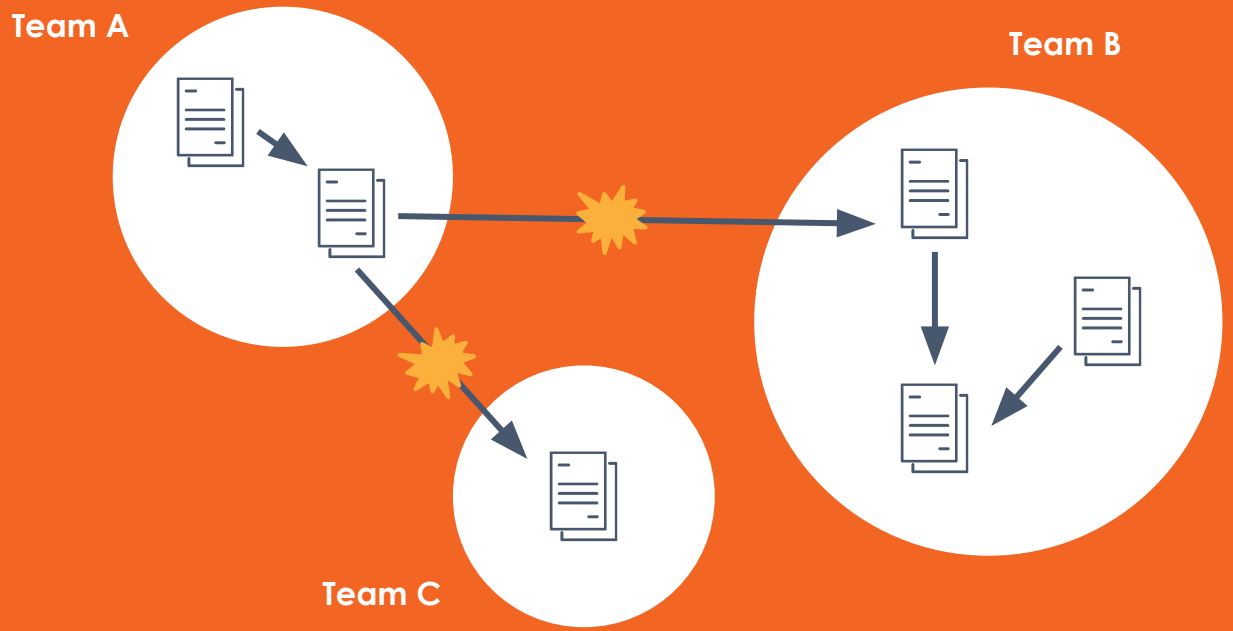
FIRSTMARK
EARLY STAGE VENTURE CAPITAL

What else changed?



The defining dilemma

Building a healthy data ecosystem



Ecosystems form around shared understanding

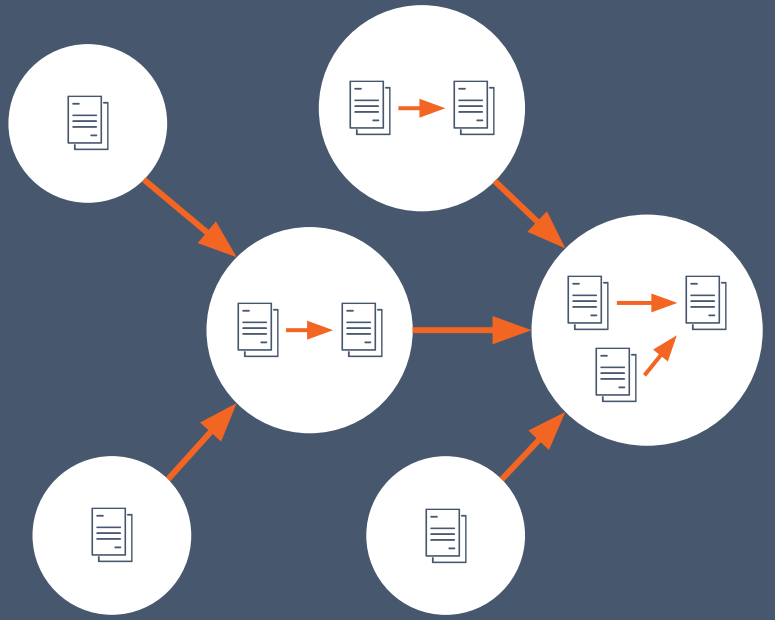
What is the data source?
What is the schema?
Who is the owner?
How often is it updated?
Where does it come from?
Who is using it?
What has changed?

What is data lineage?

Context | What | Why | How

Data lineage is the set of complex relationships between datasets and jobs in a pipeline.

- Producers & consumers of each dataset
- Inputs & outputs of each job



That’s it 

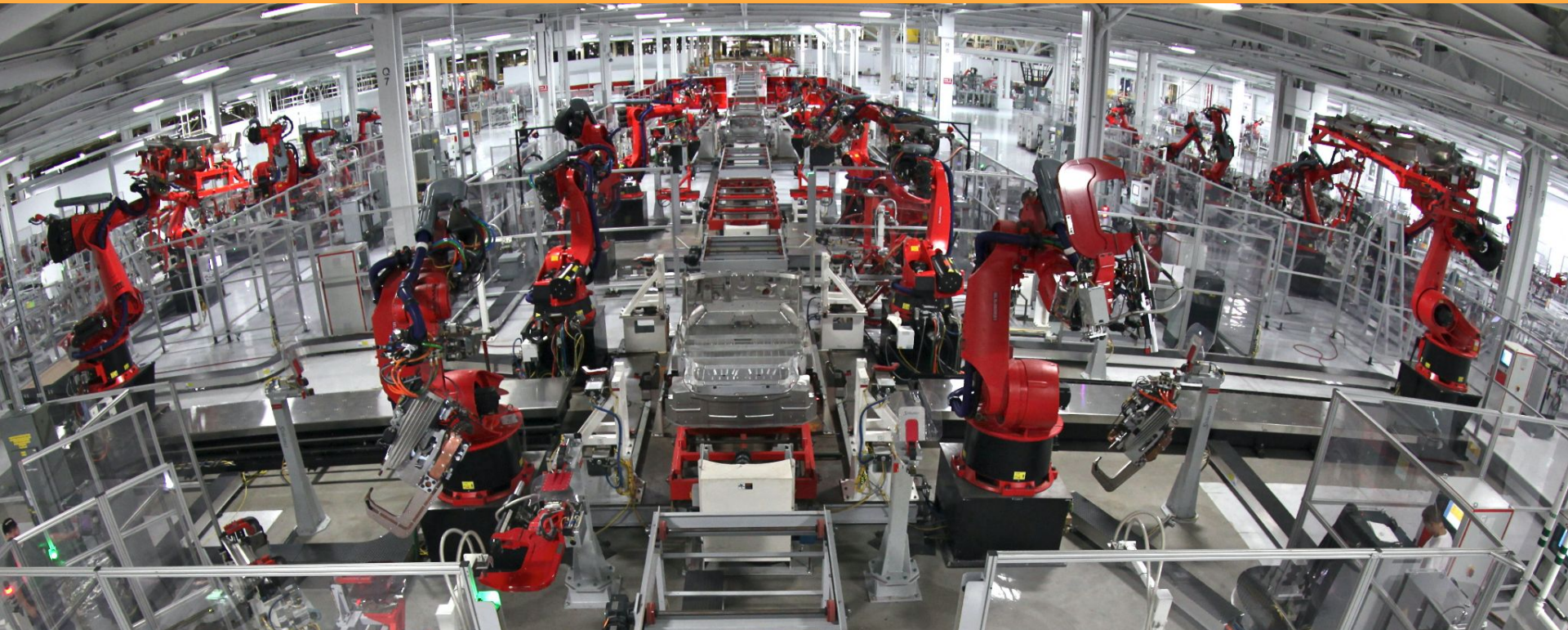
Just know everything, right?

Why do we need data lineage to succeed?

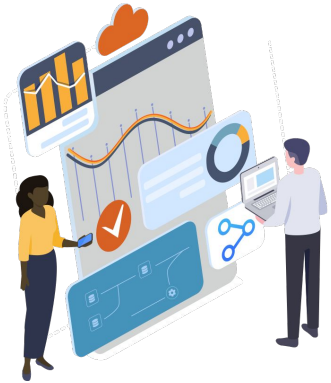
Context | What | Why | How

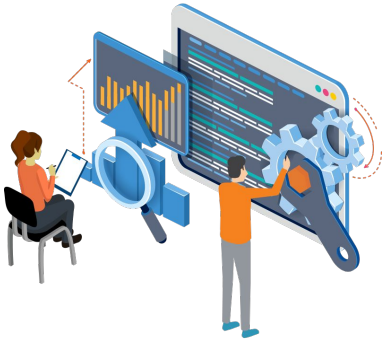
Verifying compliance

Optimizing data operations



Establishing context & language





OMG the possibilities are endless

Dependency tracing
Root cause identification
Issue prioritization
Impact mapping
Precision backfills
Anomaly detection
Change management
Historical analysis
Automated audits

Non-malicious (yet common) lineage lies

Fully-automated

Real-time

End-to-end

360° visibility

Easy

AI/ML enhanced

How can we determine data lineage?

Context | What | Why | How

EXIF has the right idea

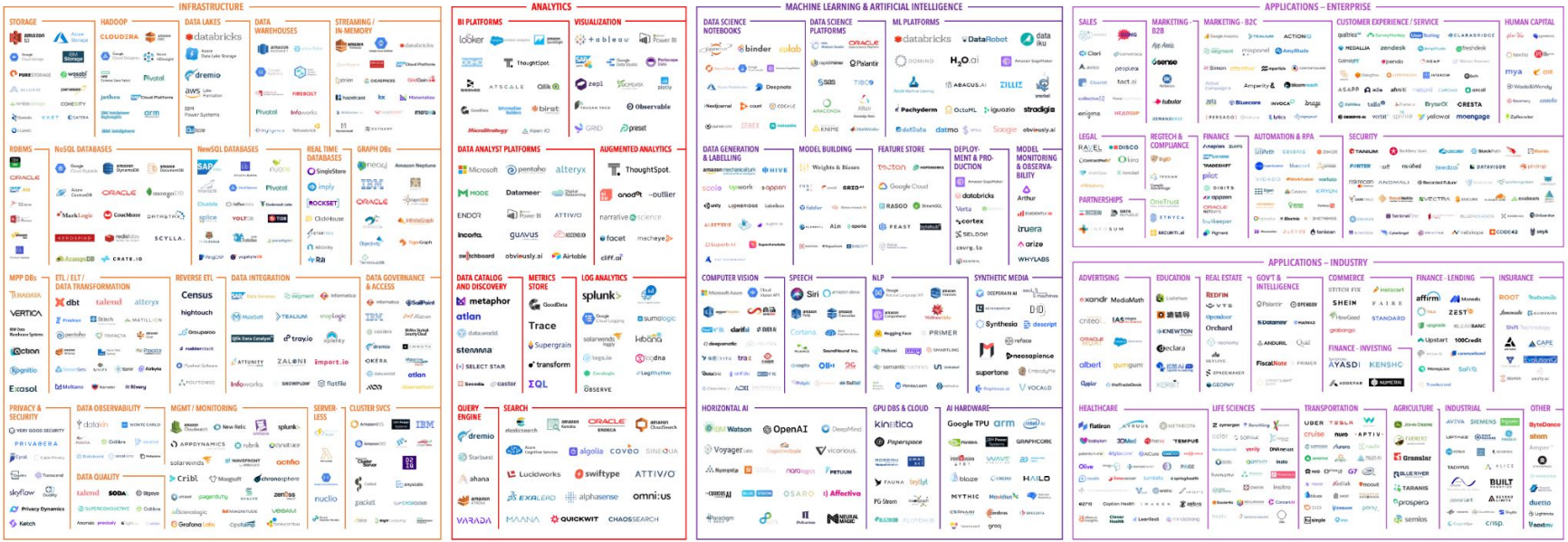
You can try to infer the date and location of an image after the fact…

…or you can capture it when the image is originally created!

That’s it 

Just collect metadata, right?

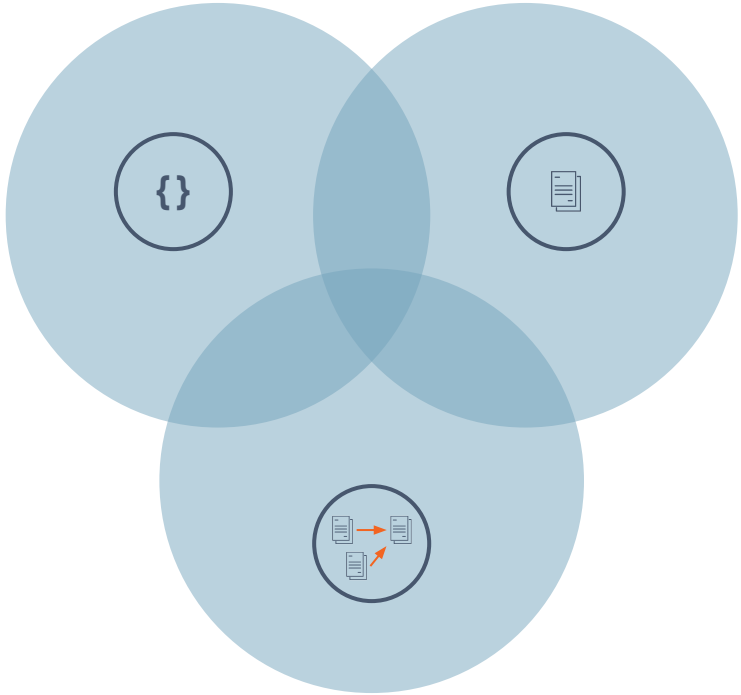
MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, AND DATA (MAD) LANDSCAPE 2021



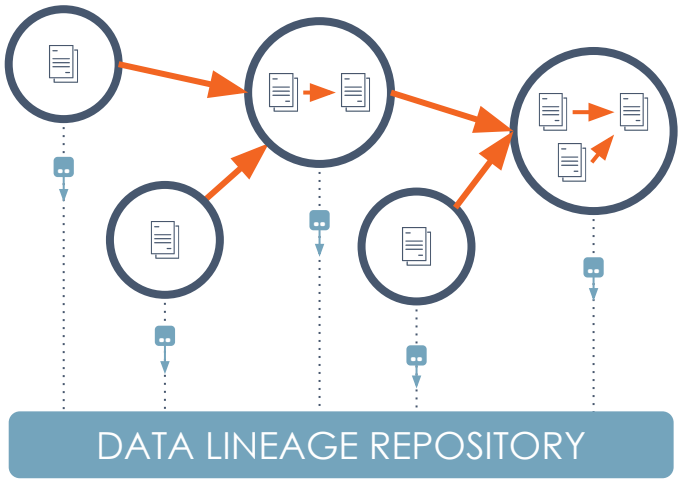
© Matt Turck (@mattturck), John Wu (@john_d_wu) & FirstMark (@firstmarkcap)

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Comparing approaches

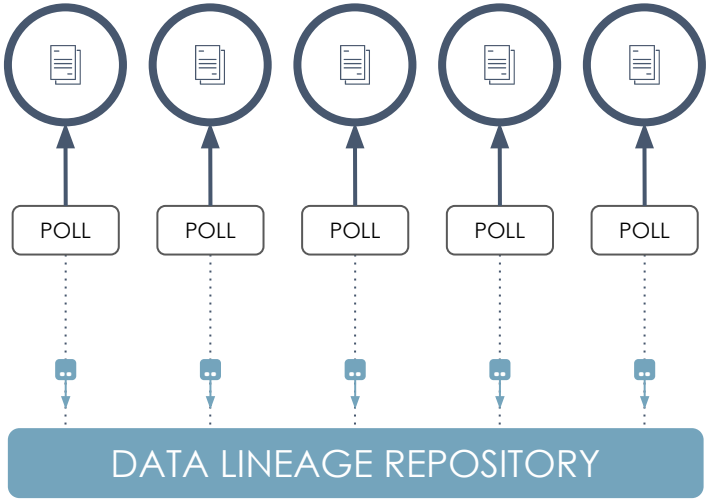


Observe the pipeline



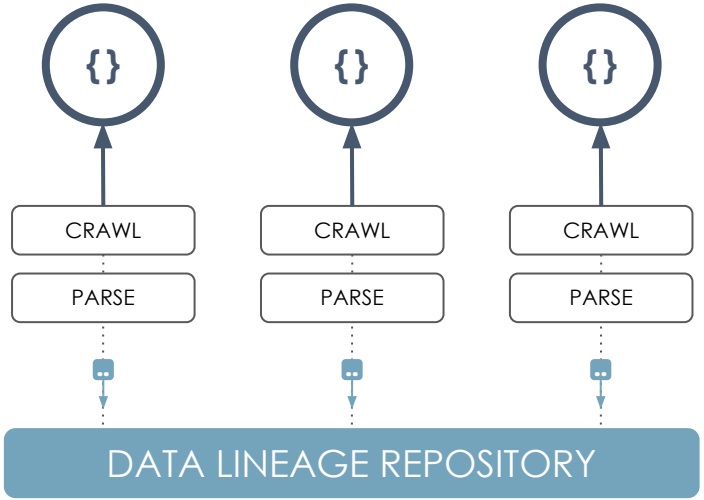
- Integrate with data orchestration systems
- As jobs run, observe the way they affect data
- Report to a lineage metadata repository

Process query / activity logs



- Integrate with data stores and warehouses
- Regularly process query logs to trace lineage
- Report to a lineage metadata repository

Analyze source code



- Integrate with source code repositories
- Look for queries and parse them for lineage
- Report to a lineage metadata repository

It’s a patchwork

A fable about community

OpenLineage

An **open standard** for the collection of lineage metadata from pipelines **as they are running**.

Standards gain momentum

The future of data lineage

Data lineage will **become standardized** and the means for gathering it will be increasingly commoditized.

Data lineage will become **increasingly multifaceted** as multiple worlds collide.

Data lineage will **work alongside orchestration** to automate operational tasks, e.g. root cause analysis and backfills.

Thanks :)