Data Quality and Observability with Airflow

3.0

Chirag Tailor
Ipsa Trivedi

Agenda

- About Us
- Our Use Case
- Observability Options
- Our Solution
- Where Our Solution Shines
- Next Steps

About Us

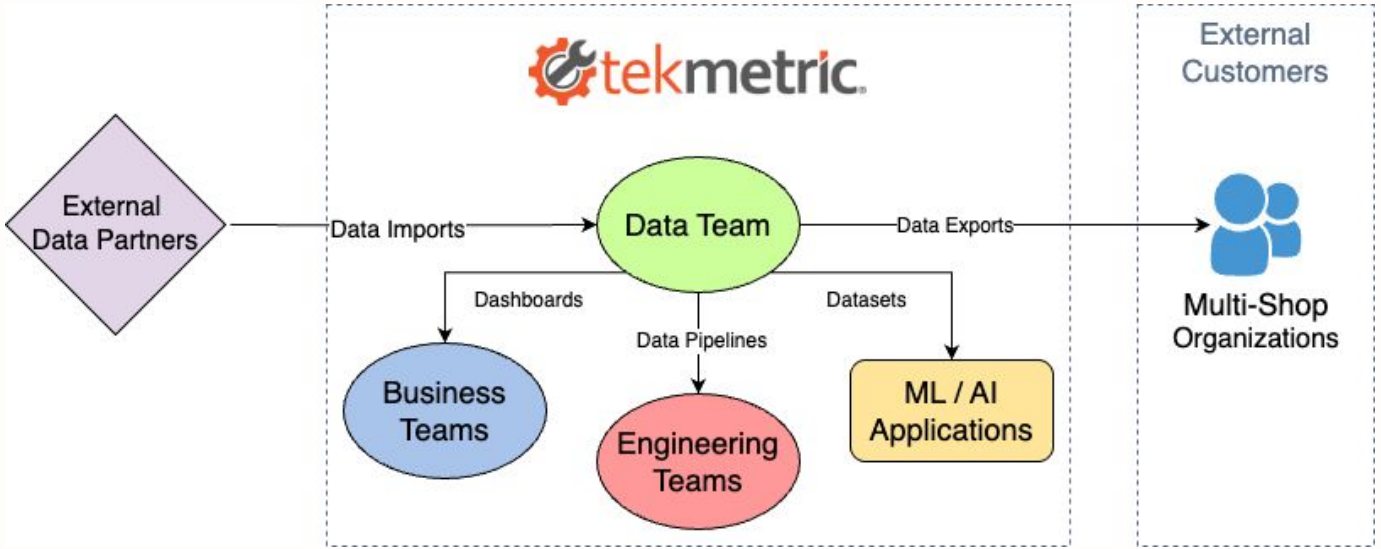
About Us

Tekmetric

- Initial product launch in 2017
- Largest Cloud-based Automotive Shop Management System in the US
- Our platform supports over 11,000 shops and provides an array of valuable products:
  - Estimate Building, Parts & Inventory Management
  - Payments
  - CRM
  - Marketing
  - Reporting
- This year, we have already processed 13.2M Repair Orders for a total of $8.4B.

Our Use Case

Use Case



Problem Statement

- Ensure data quality for machine learning training pipelines
- Strict data veracity standards to build trust with external partners

Observability Options

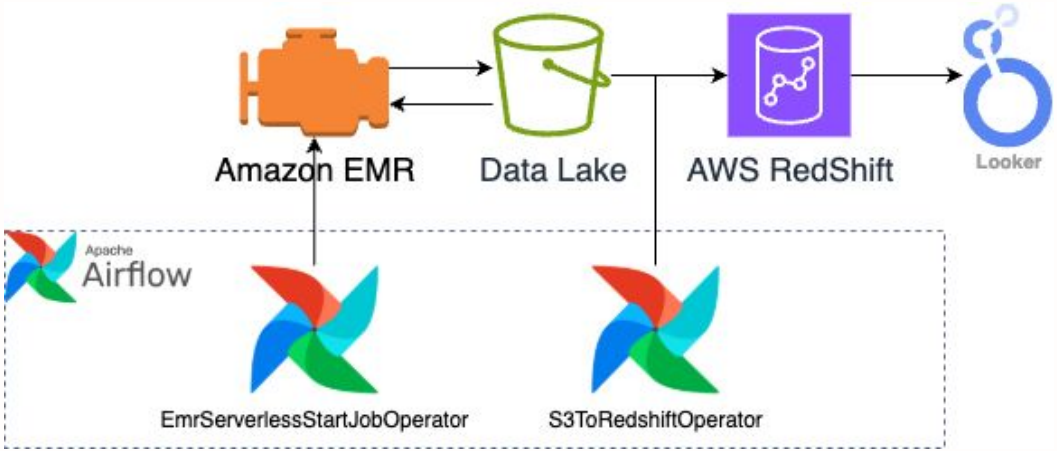
Our Options

1. Airflow + Looker

1. DQ Tools (Great Expectations, Colibra, Soda, etc.)

1. Prometheus & Grafana

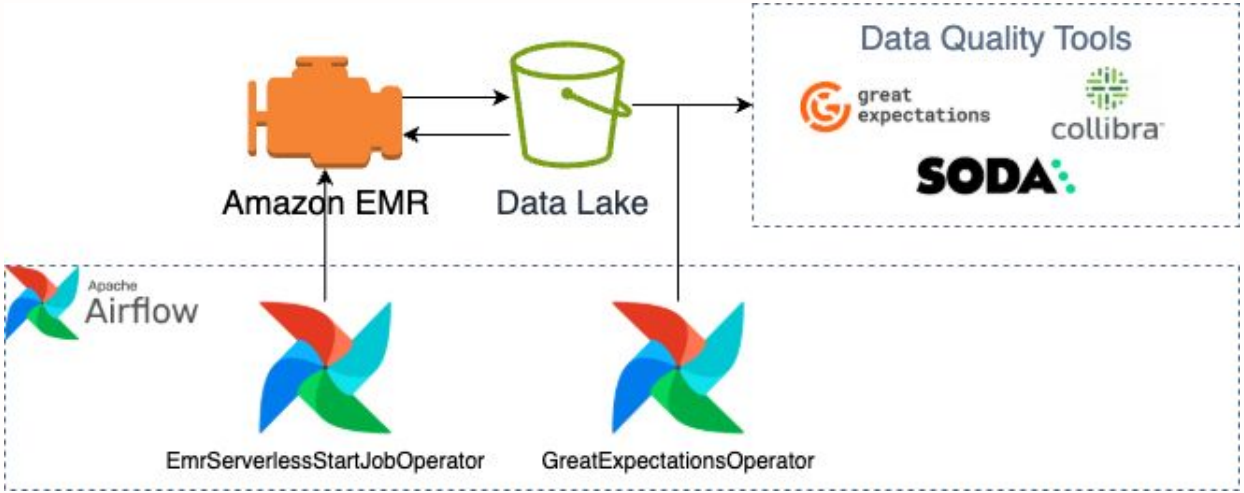
Our Options - Airflow & Looker



Does it fit our Use Cases?

- x Inability to view data as time series metrics
- x Inability to group visualizations by labels
- ✓ Inability to identify trends or anomalies in metrics
- x Inability to assist with data profiling

Our Options - Various DQ Tools



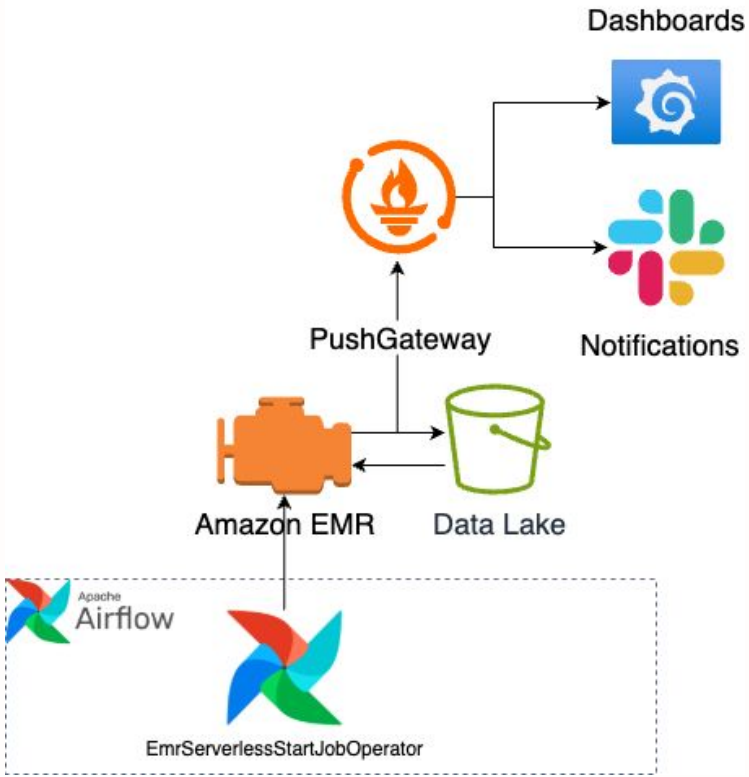
Does it fit our Use Cases?

- x Inability to view data as time series metrics
- x Inability to group metrics by labels
- ✓ Ability to identify trends or anomalies in metrics
- ✓ Ability to assist with data profiling

Our Options - Prometheus & Grafana

Does it fit our Use Cases?

- ✓ Ability to view data as time series metrics
- ✓ Ability to group visualizations by labels
- ✓ Ability to identify trends or anomalies in metrics
- x Inability to assist with data profiling

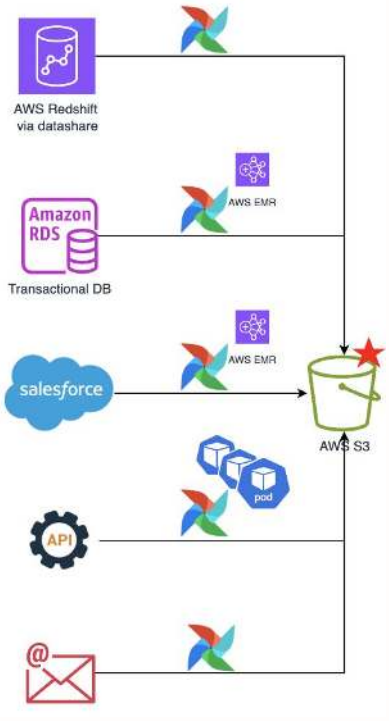


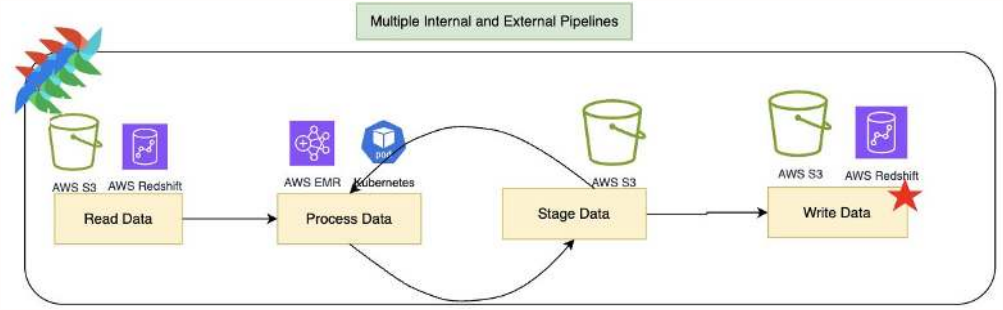
Our Solution

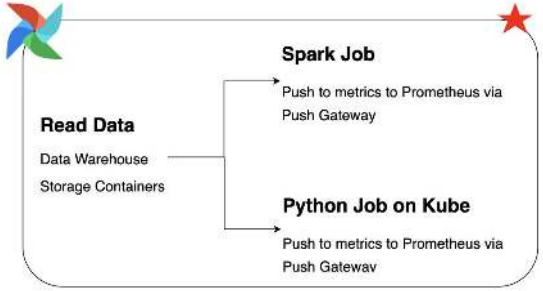
Why Prometheus & Grafana?

1. Custom Metric Python Library (PushGateway)
2. Slack Alerting
3. Powerful Dashboard Capabilities
4. Leverage Existing Observability Stack
5. Scales well with our datasets

Design

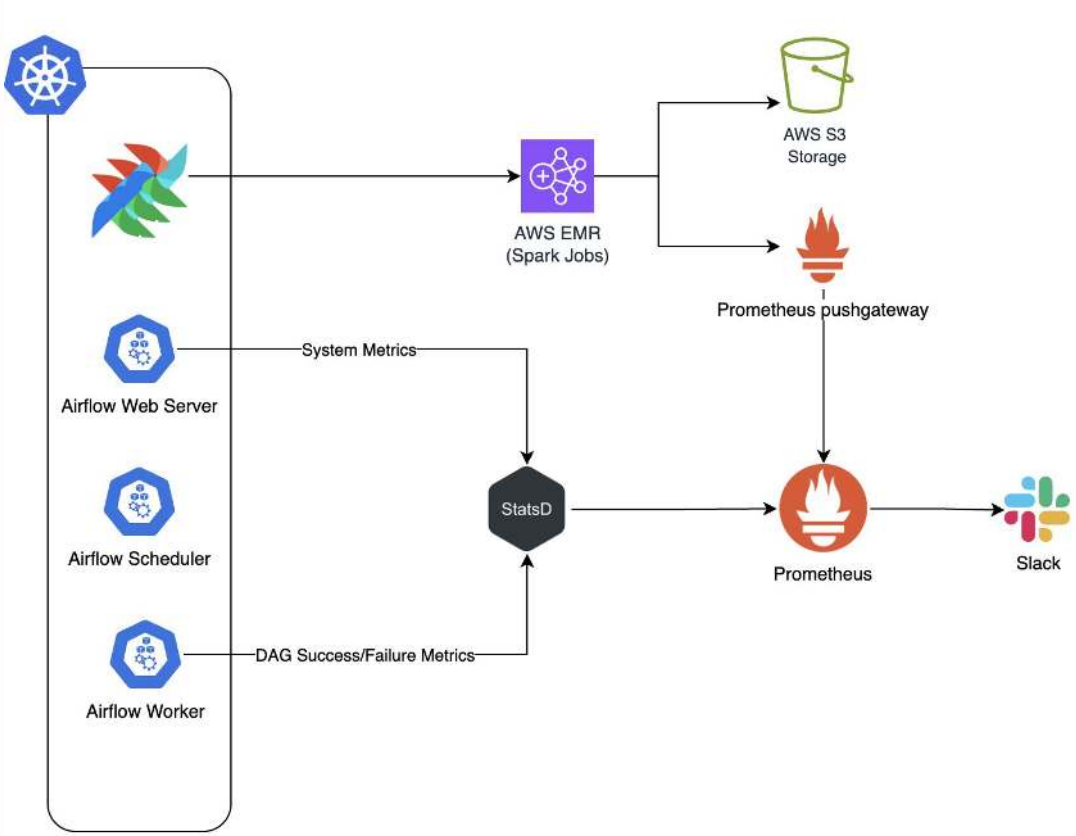

AWS Redshift
via datashare
AWS EMR
Amazon
RDS
Transactional DB
AWS EMR
salesforce
AWS S3
pod
API
@


Multiple Internal and External Pipelines
AWS S3
AWS Redshift
Read Data
AWS EMR
Kubernetes
Process Data
AWS S3
Stage Data
AWS S3
AWS Redshift
Write Data

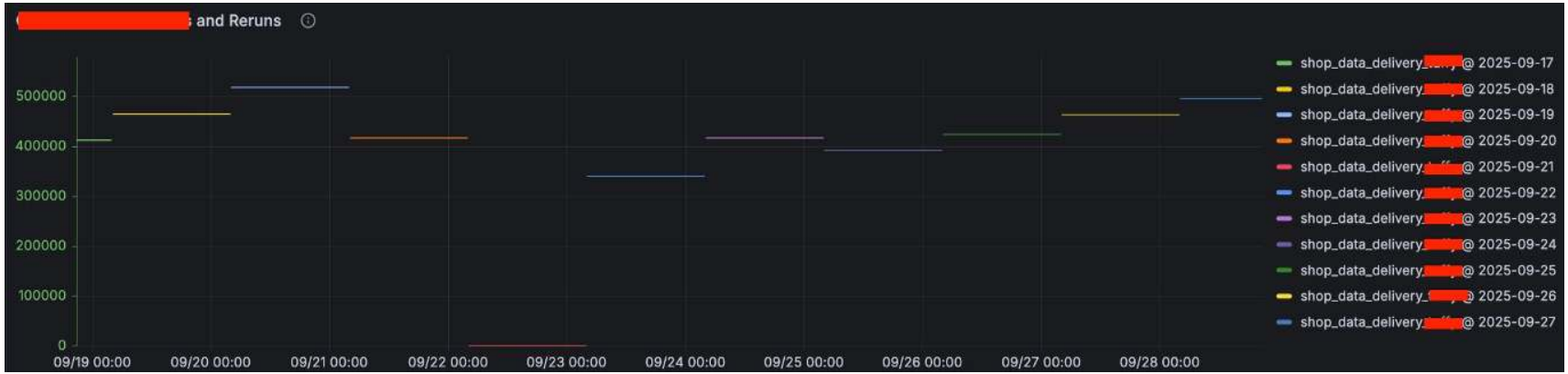

Spark Job
Push to metrics to Prometheus via
Push Gateway
Read Data
Data Warehouse
Storage Containers
Python Job on Kube
Push to metrics to Prometheus via
Push Gateway

Looker
Machine Learning
AWS Bedrock
AWS S3
Email
Cloud Storages
On Prem Servers

Architecture

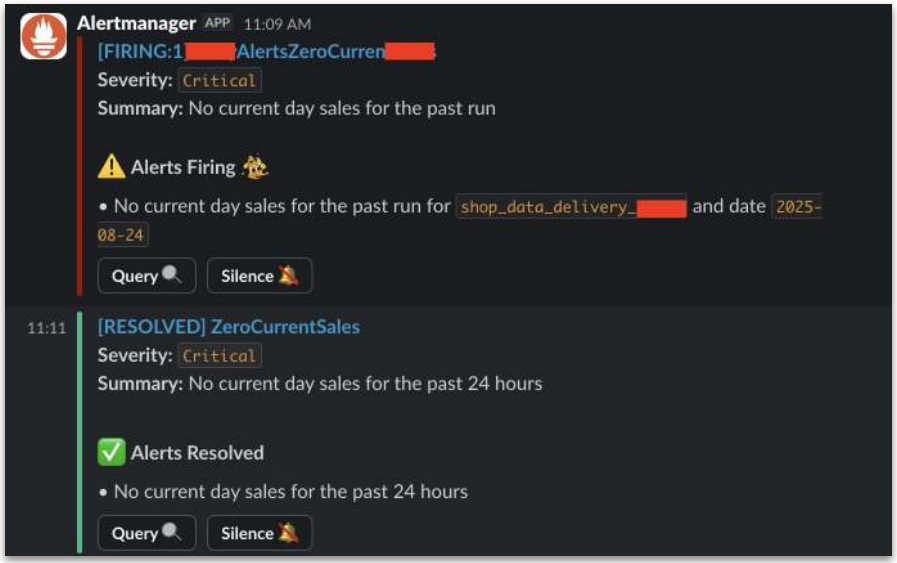


Data Obs: Dashboards

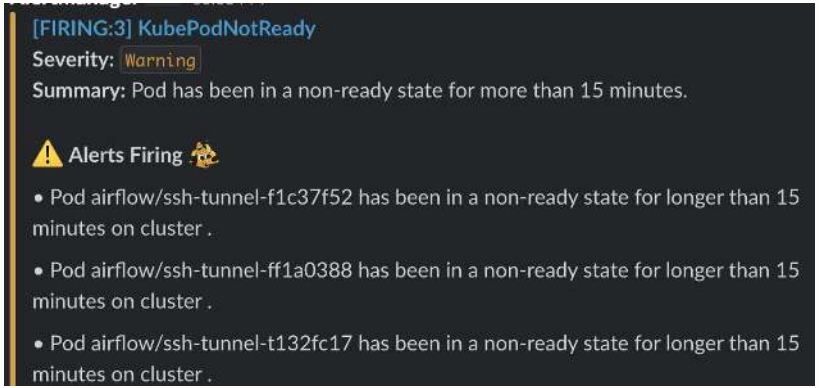


Data Obs: Alerts

Highlighting DQ issues and resolution alerts



DAG Failure Alerts



Where Our Solution Shines

Strengths

Advantages

- Most companies already have Prometheus setup for system observability
- Seamless integration with Airflow
- Visualization with Grafana
- System health and data health -> **one stop shop for observability**
- Low cost option due to it being Open Source Software (OSS)

Next Steps

Considering Data Quality Tools

When...

- As checks grow complex (cross-dataset, lineage-aware, profiling, auto-discovery)
- Manual rule-writing becomes a bottleneck
- Real-time metric feedback and automated pipeline reruns/reconciliation are needed
- Long-term data retention is required

How..

- Use DQ tools (e.g., Great Expectations) integrated with Airflow.
- Adopt a hybrid model: DQ tool for profiling/lineage, Prometheus for ops & metrics.

Questions?