



Data Quality and Observability with Airflow

3.0

A large, stylized blue number "3.0" is positioned on the right side of the slide. The "3" is a tall, rounded digit, and the ".0" is a smaller, thinner digit. The entire number is set against a black background.

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

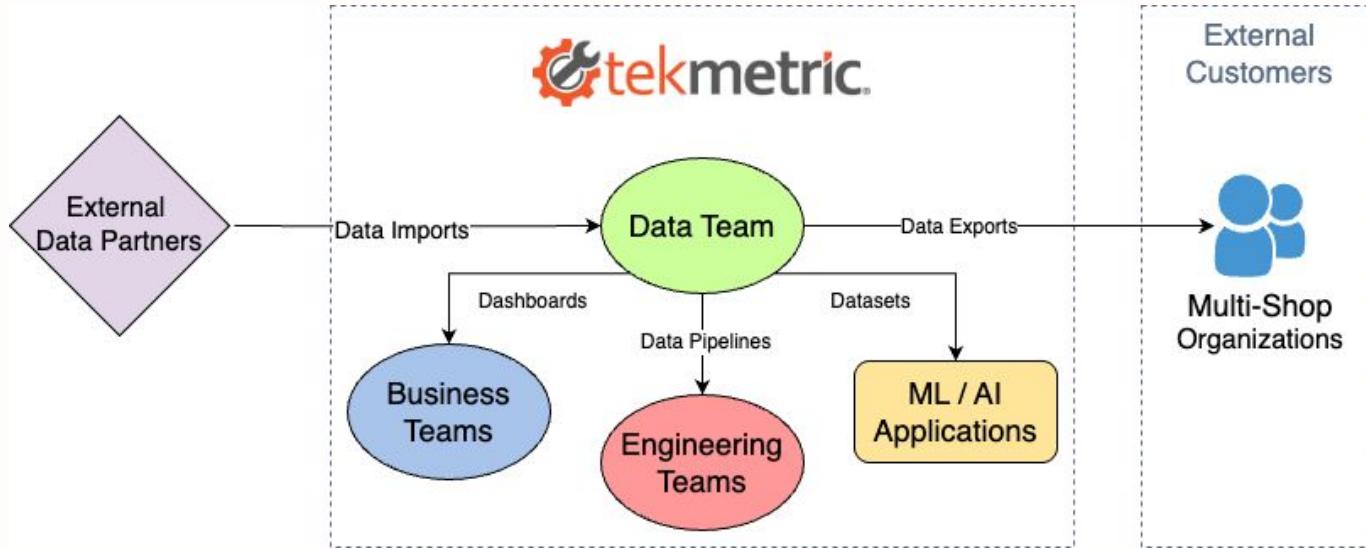
- Initial product launch in 2017
- Largest Cloud-based Automotive Shop Management System in the US
- Our platform supports over 11,000 shops and provides an array of valuable products:
 - Estimate Building, Parts & Inventory Management
 - Payments
 - CRM
 - Marketing
 - Reporting
- This year, we have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



Our Use Case



Use Case

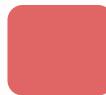


Problem Statement

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



Observability Options



Our Options

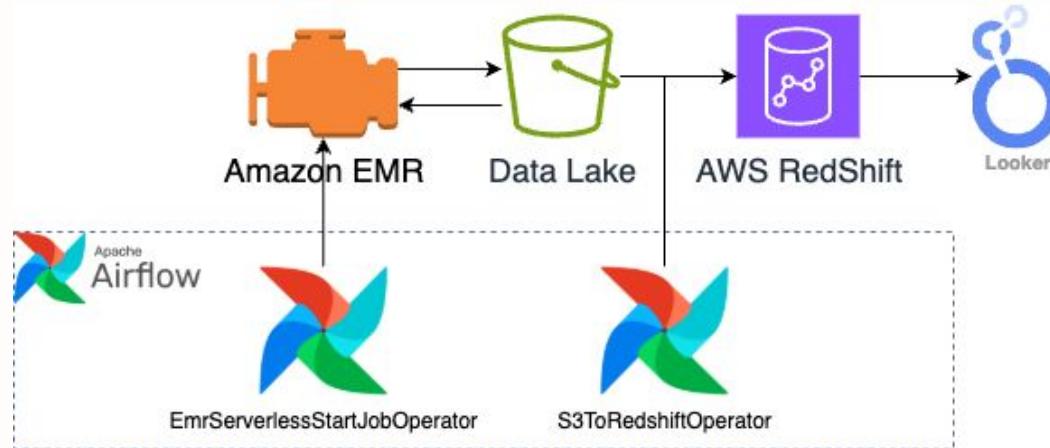
1. Airflow + Looker

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

1. Prometheus & Grafana



Our Options - Airflow & Looker

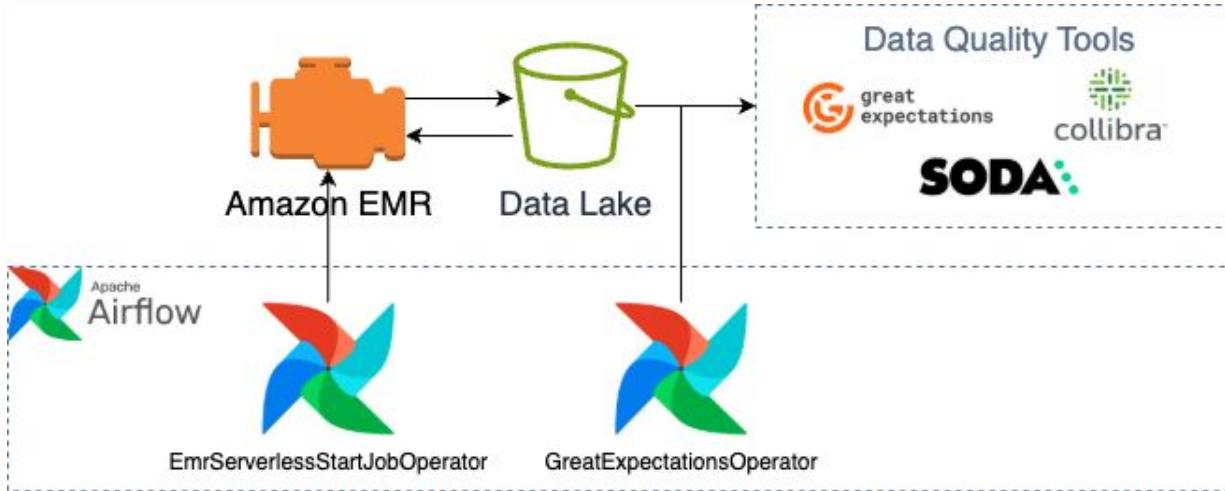


Does it fit our Use Cases?

- 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?

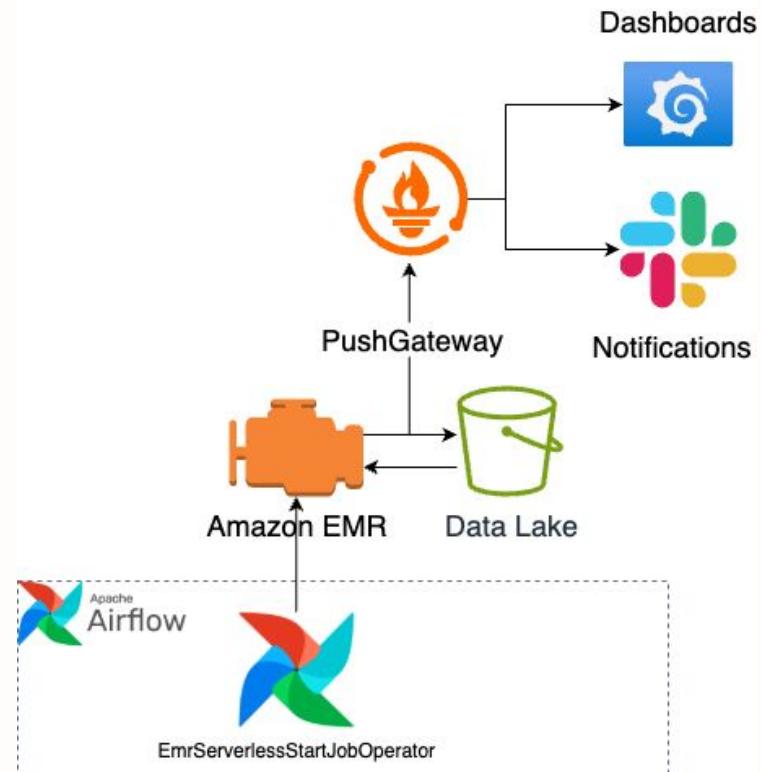
- x Inability to view data as time series metrics
- x Inability to group metrics by labels
- ✓ Ability to identify trends or anomalies in metrics
- ✓ Ability to assist with data profiling



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
- ✗ 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

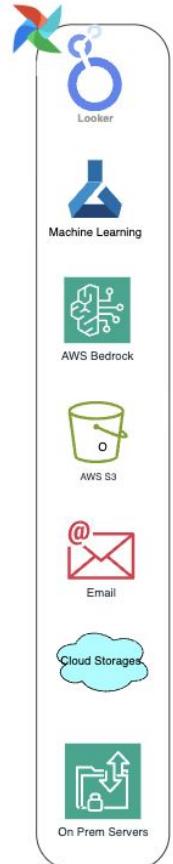
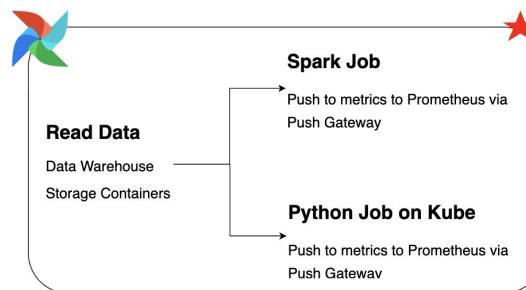
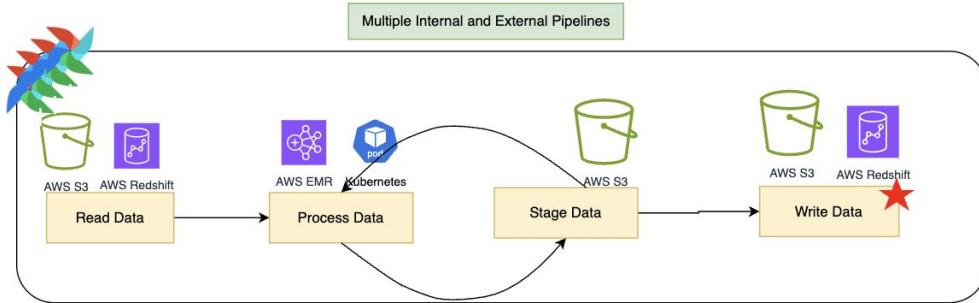
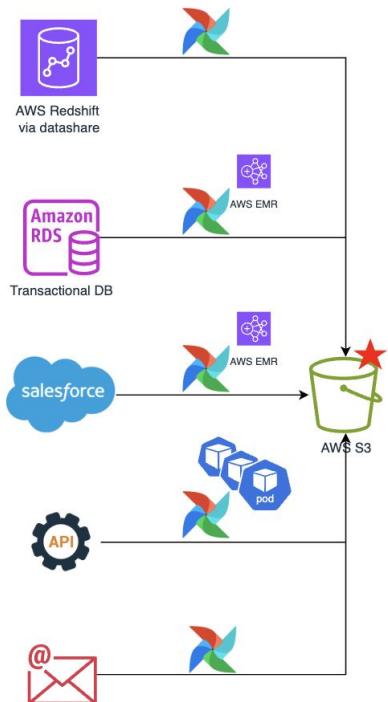


Prometheus

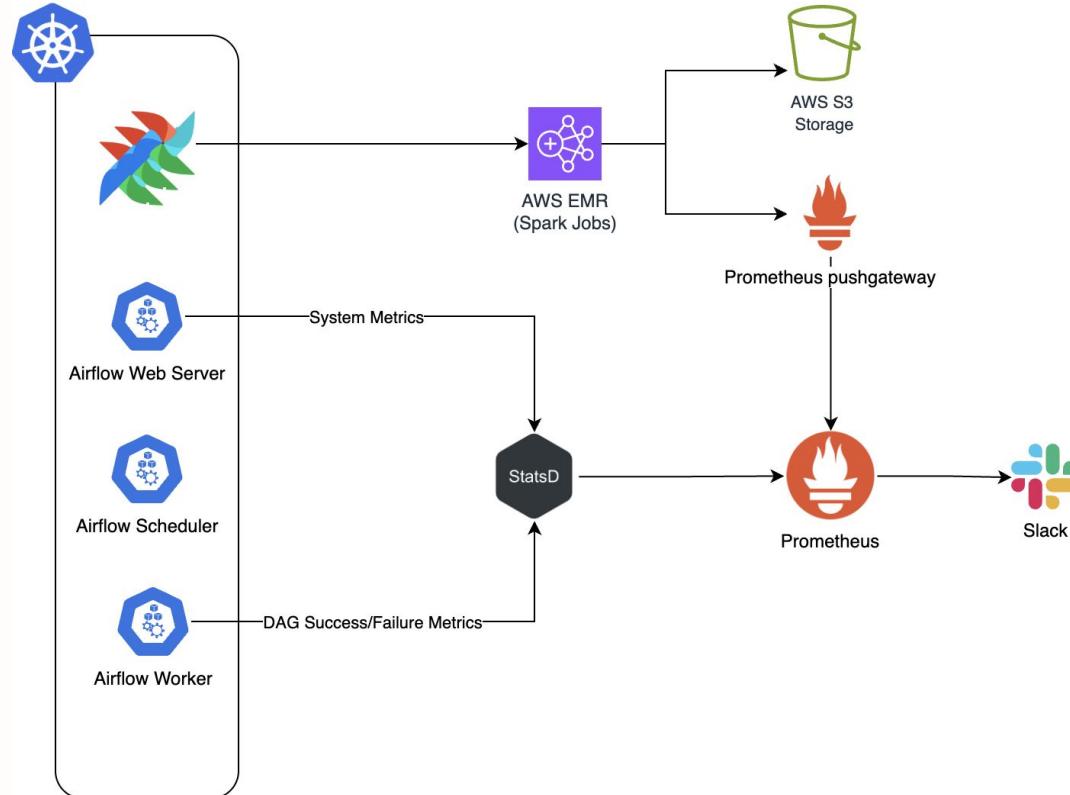


Grafana

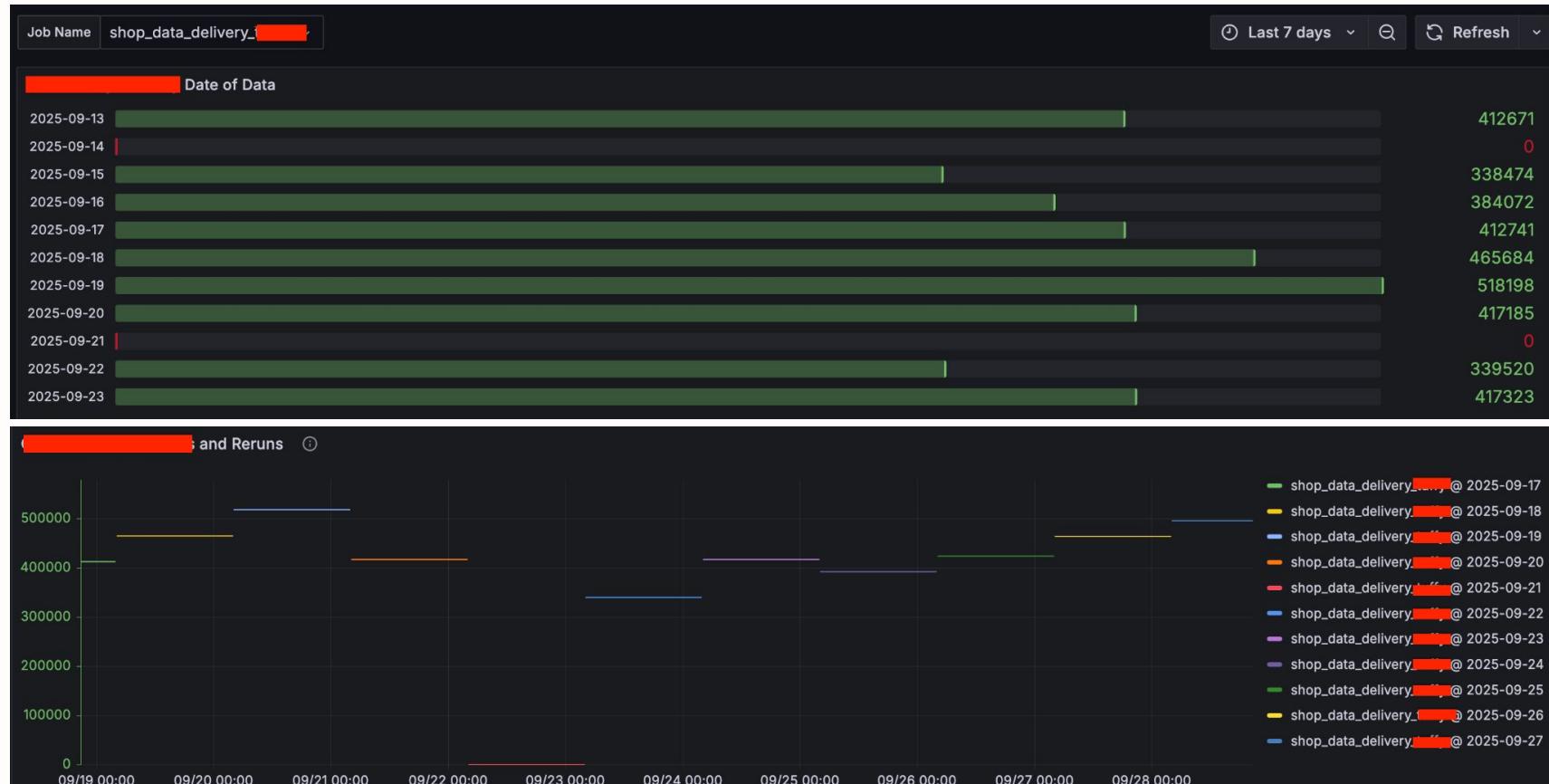
Design



Architecture



Data Obs: Dashboards



Data Obs: Alerts

Highlighting DQ issues and resolution alerts

Alertmanager APP 11:09 AM
[FIRING:1] AlertsZeroCurrentSales
Severity: Critical
Summary: No current day sales for the past run

⚠️ Alerts Firing 🚨
• No current day sales for the past run for shop_data_delivery_t [REDACTED] and date 2025-08-24

Query 🔎 **Silence 🕊**

11:11 [RESOLVED] ZeroCurrentSales
Severity: Critical
Summary: No current day sales for the past 24 hours

✓ Alerts Resolved
• No current day sales for the past 24 hours

Query 🔎 **Silence 🕊**

DAG Failure Alerts

Alertmanager APP May 1st at 12:50 PM
[FIRING:1] AirflowDagrunFailed
Severity: Critical
Summary: Airflow DagRun Failure

⚠️ Alerts Firing 🚨
• The Airflow DAG, with ID shop_data_delivery_t [REDACTED], has failed.

Query 🔎 **Silence 🕊**

[FIRING:3] KubePodNotReady
Severity: Warning
Summary: Pod has been in a non-ready state for more than 15 minutes.

⚠️ Alerts Firing 🚨
• Pod airflow/ssh-tunnel-f1c37f52 has been in a non-ready state for longer than 15 minutes on cluster .
• Pod airflow/ssh-tunnel-ff1a0388 has been in a non-ready state for longer than 15 minutes on cluster .
• Pod airflow/ssh-tunnel-t132fc17 has been in a non-ready state for longer than 15 minutes on cluster .



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?