



# Data Quality at Atlassian



YASH SHARMA | SENIOR DATA ENGINEER | ATLASSIAN



 Bitbucket

 Jira

 Confluence

 Sourcetree

 Statuspage

 Trello

# And our Values



**Open company,  
no bullshit**



**Build with heart  
and balance**



**Don't #@!%  
the customer**



**Play, as a team**

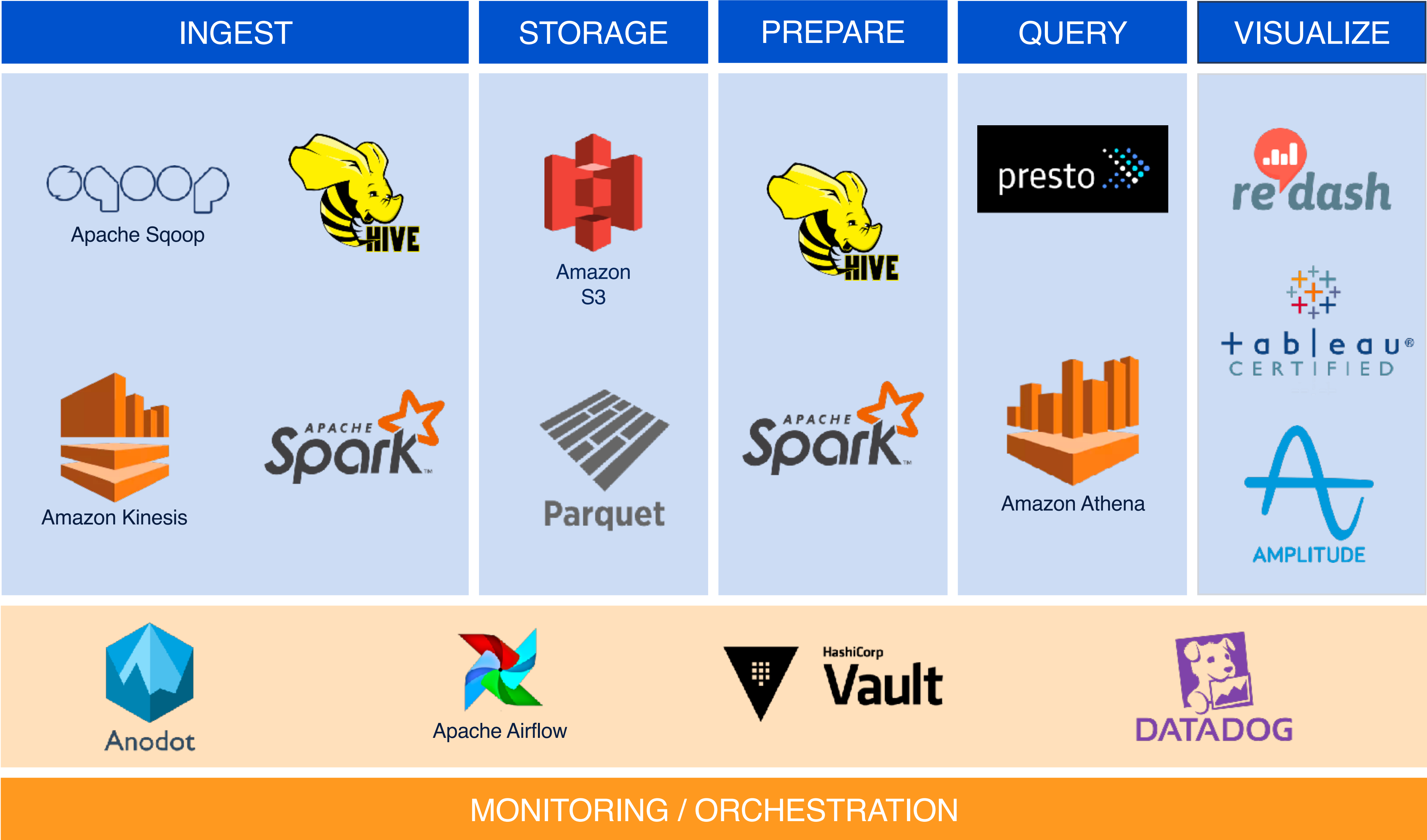


**Be the change you seek**



# Data Engineering





# Some History



# Self-served data culture

# Self-served data culture



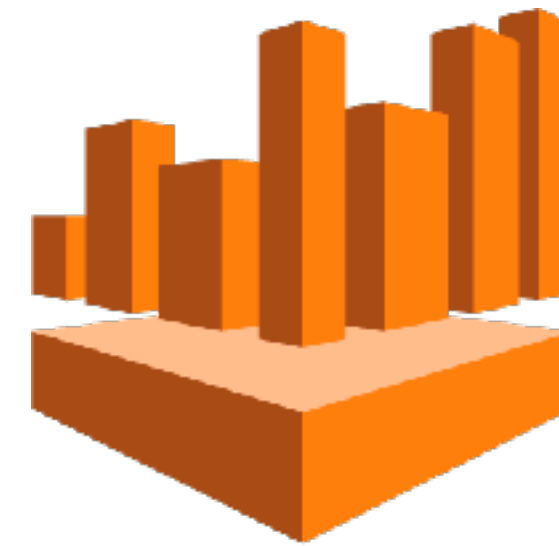
**1800 Monthly Users**

Using the analytics platform each month



**260k Monthly Queries**

Executed on our analytical engines



**AWS Athena**

160 schemas and  
5000+ Tables



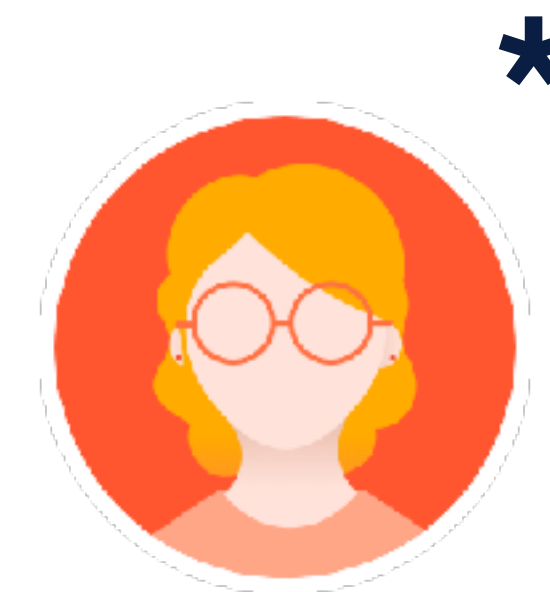
**Reporting**

35,000 Redash Reports  
2000+ Redash Dashboards  
500+ Tableau Workbooks

# What about data quality



**10 Analyst**

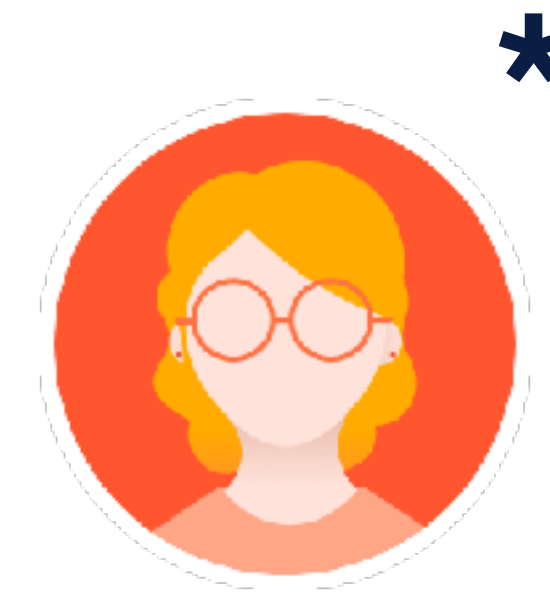


**1 Data Engineer**

# What about data quality



**10 Analyst**



**1 Data Engineer**

\* fake data

# The Plan

Goals

 **Self serve data quality**

**Self served**

 **Ownership**

Usable

Actionable

**Goals**

---

Self Served

**Usable**

Actionable

- + Ease of use**
- + Decoupled from pipeline**
- + Scheduled and Ad-hoc**
- + Integrations**

Goals

---

Self Served

Usable

Actionable

 **Alert on failure**

 **Escalation**

 **Save check metadata**





# Introducing Yoda

# Yoda : Quick Demo





# Data Issue Lineage



66



## ANALYTICS



## Debugger




 Refresh

+ Create

Archived

ALL 



	Schedule	Name	Author	Owner	Status	Last Status Change	
<input type="checkbox"/>	 10 01 * * *	Nps event cloud	ysharma	zone_ai	SUCCESS	4 months ago	...
<input type="checkbox"/>	 10 * * * *	Grow events	ysharma	zone_ai	RUNNING	4 months ago	...
<input type="checkbox"/>	 10 01 * * *	Cloud event summary	ysharma	zone_ai	SUCCESS	4 months ago	...

< 1 >





## ANALYTICS



## Debugger




 Refresh

+ Create

Archived

ALL



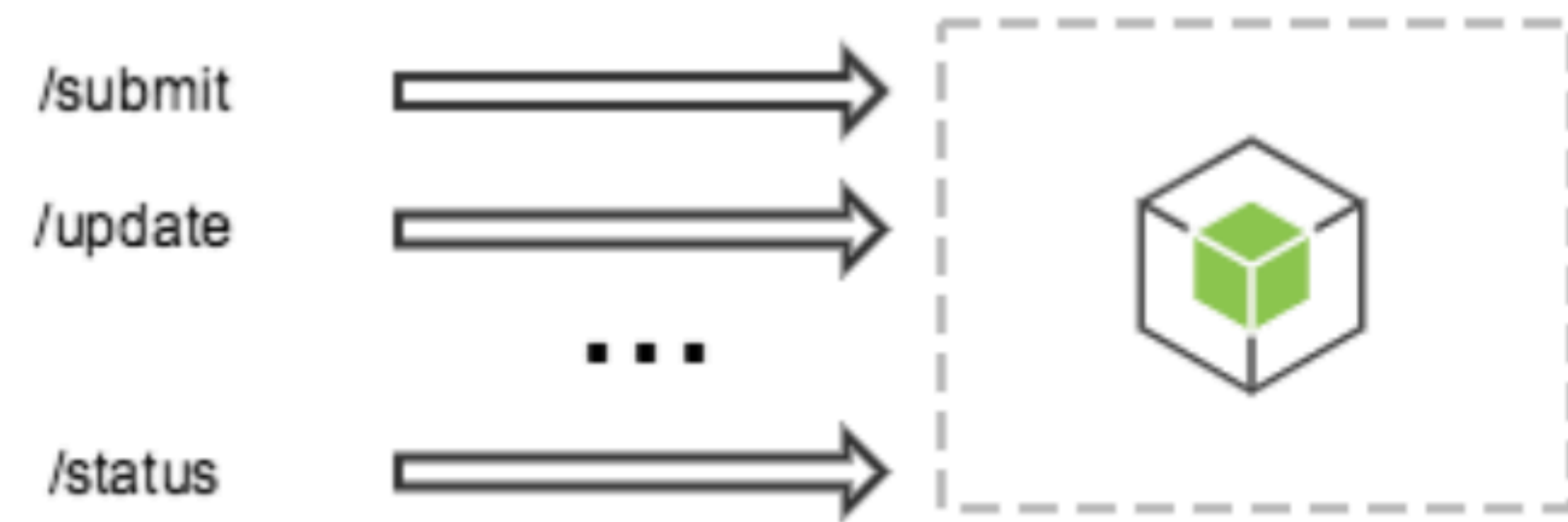
	Schedule	Name	Author	Owner	Status	Last Status Change	
<input type="checkbox"/>	 10 01 * * *	Nps event cloud	ysharma	zone_ai	SUCCESS	4 months ago	...
<input type="checkbox"/>	 10 * * * *	Grow events	ysharma	zone_ai	RUNNING	4 months ago	...
<input type="checkbox"/>	 10 01 * * *	Cloud event summary	ysharma	zone_ai	SUCCESS	4 months ago	...

< 1 >

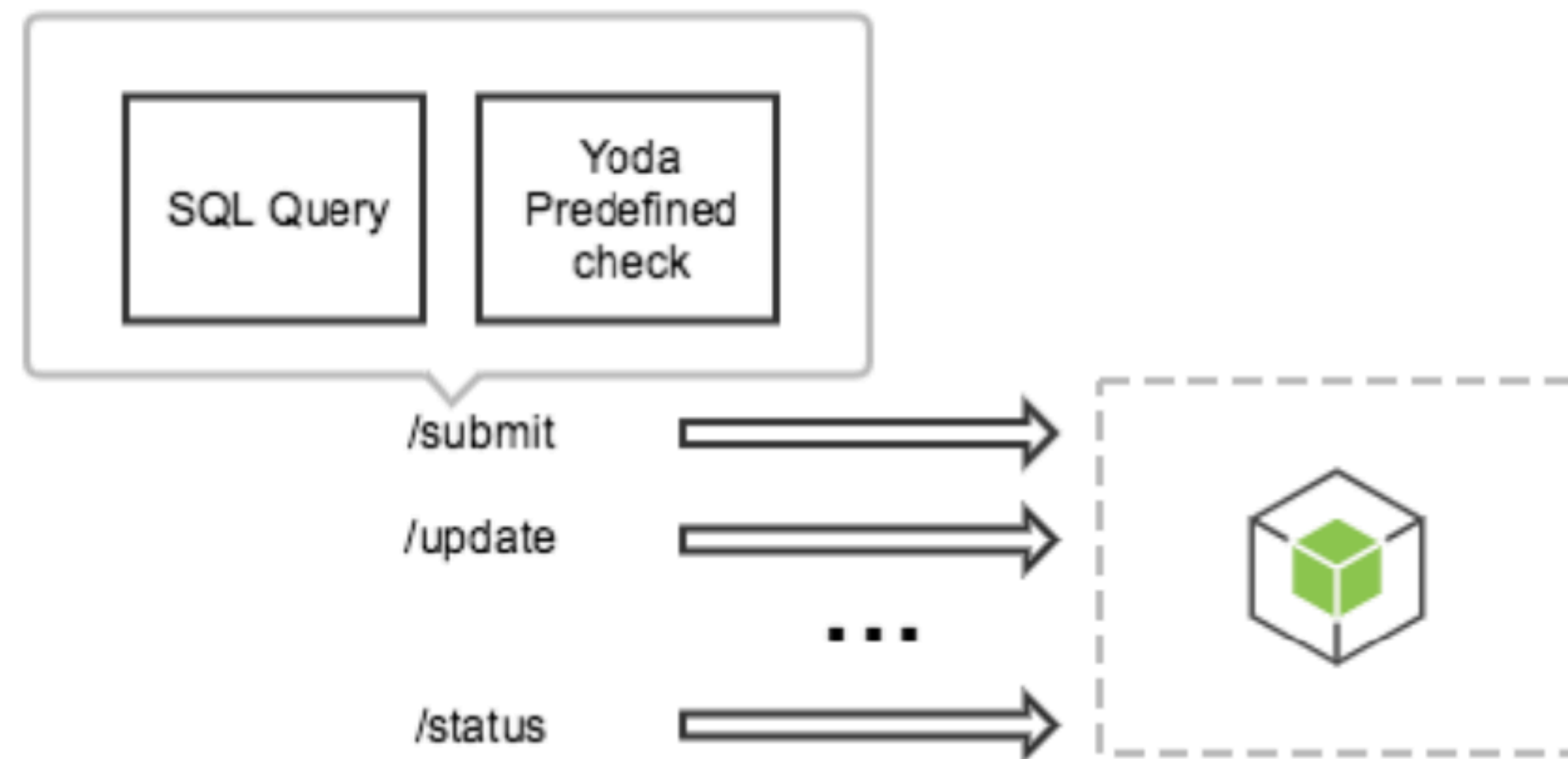
# The Implementation



# Micro service for quality checks



# Scheduled vs AdHoc and SQL vs Config

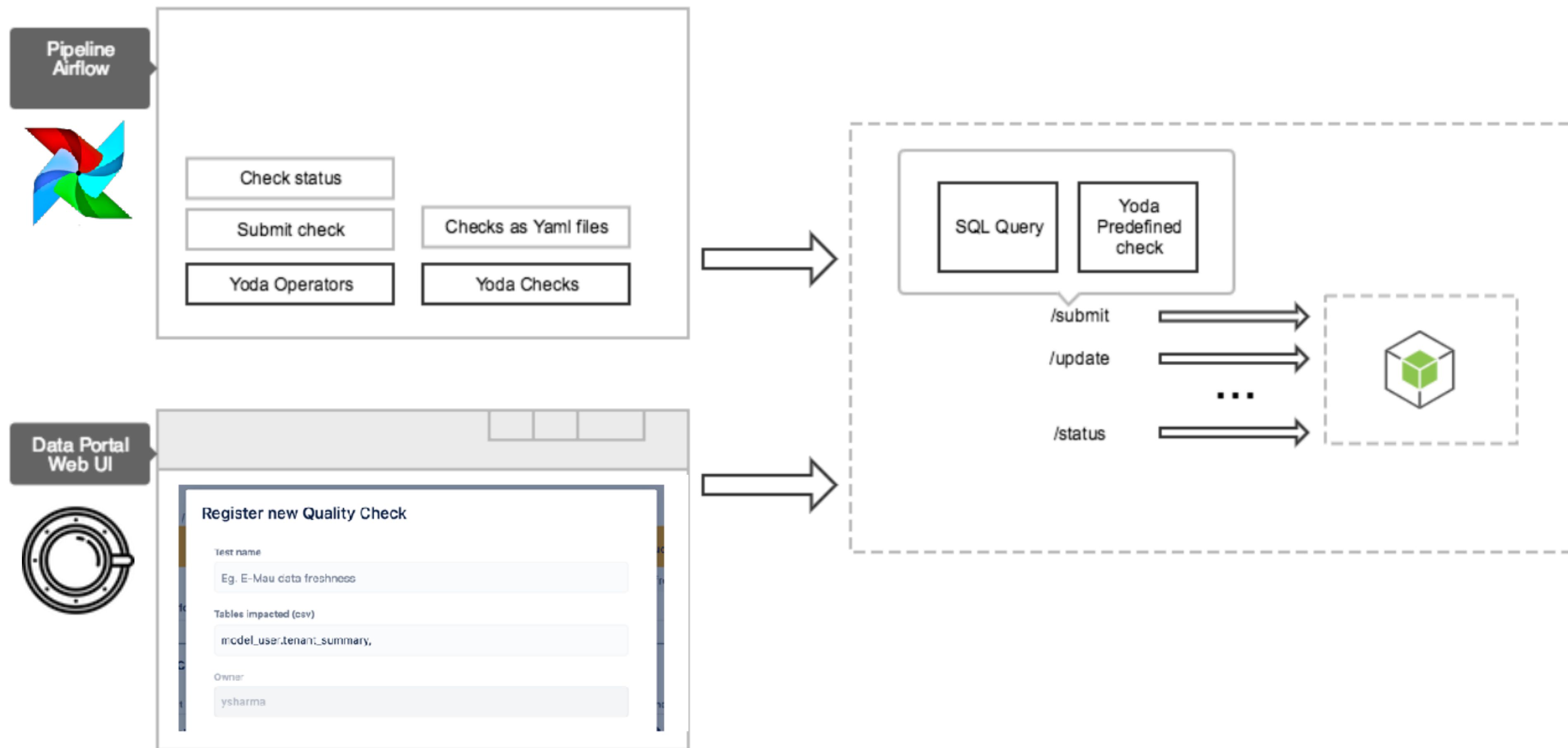


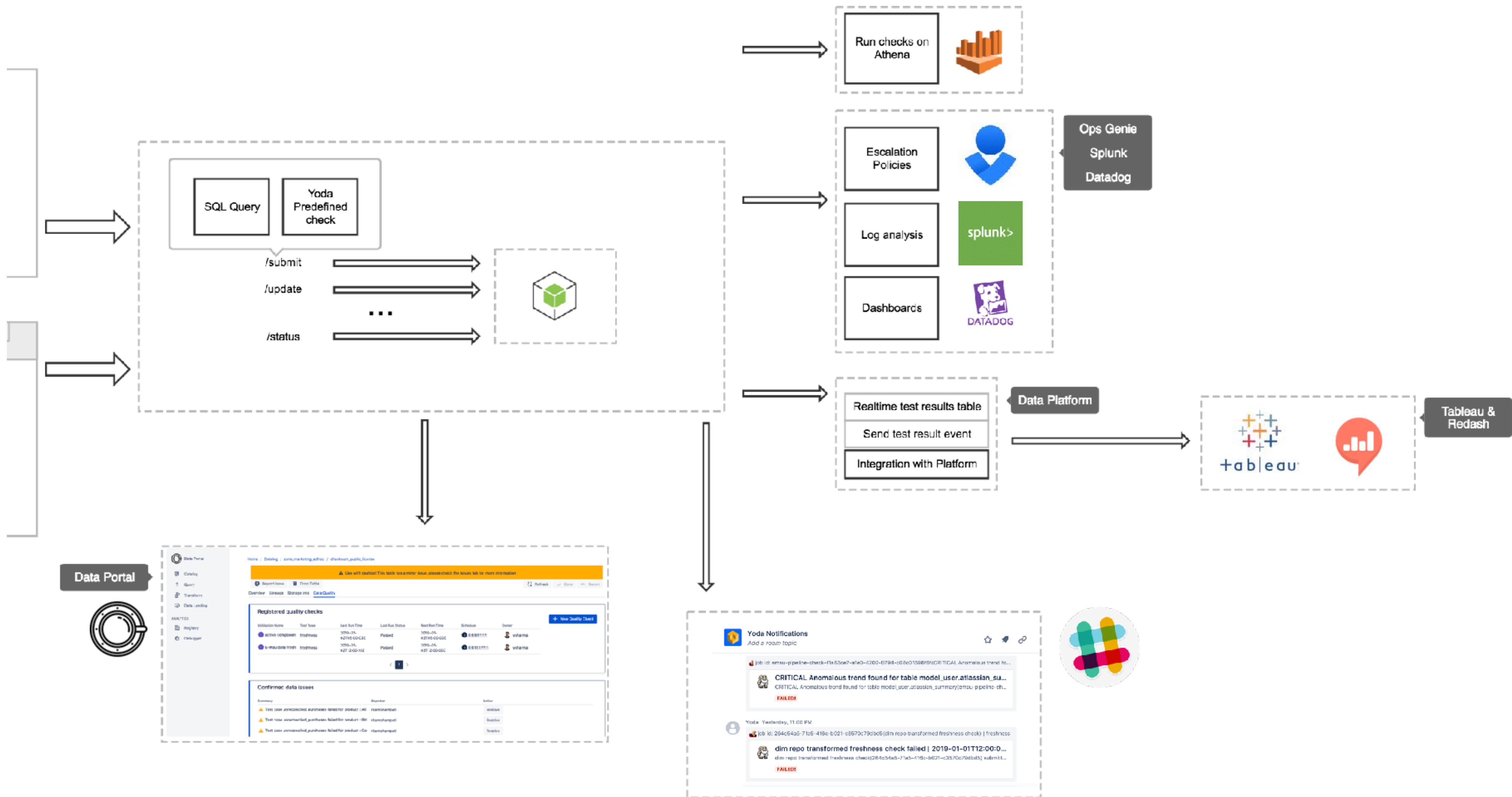
# SQL check vs Predefined check

```
WITH check_logic as (  
  )  
SELECT  
  'confluence' AS product,  
  'cloud' AS platform,  
  <False for failure> AS result,  
  <JSON Metadata> AS additional_data  
FROM check_logic
```

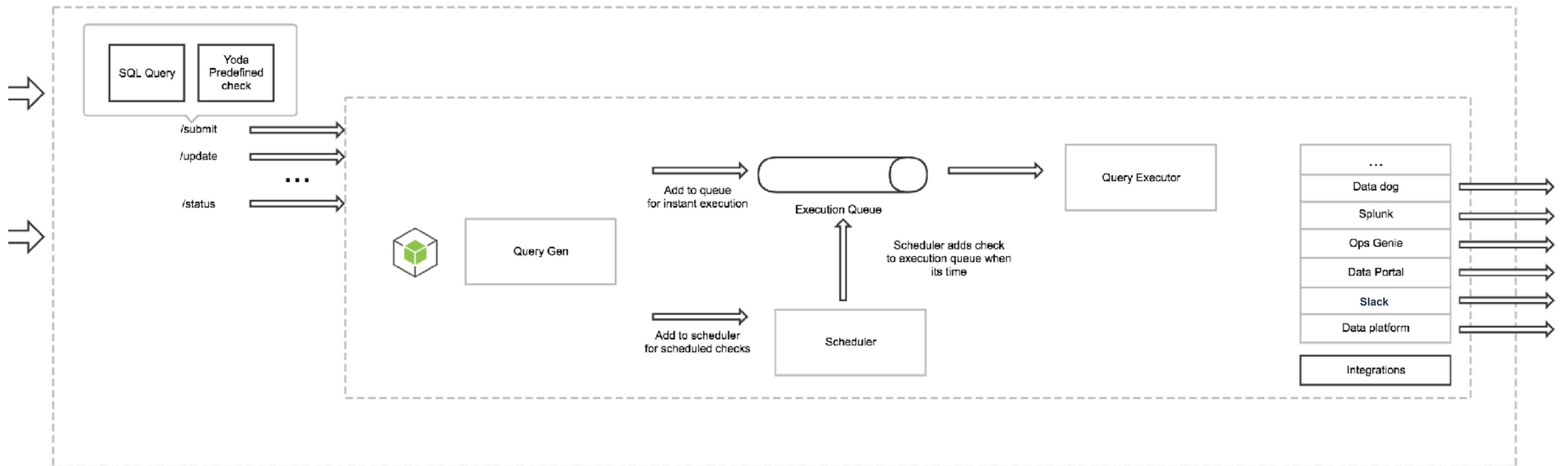
```
predefined_check:  
  check_type: column_should_have_distinct_values  
  date_column: day  
  table_name: model_user.product_summary  
  composite_columns:  
    - day  
    - product  
    - metric
```

# Integrations

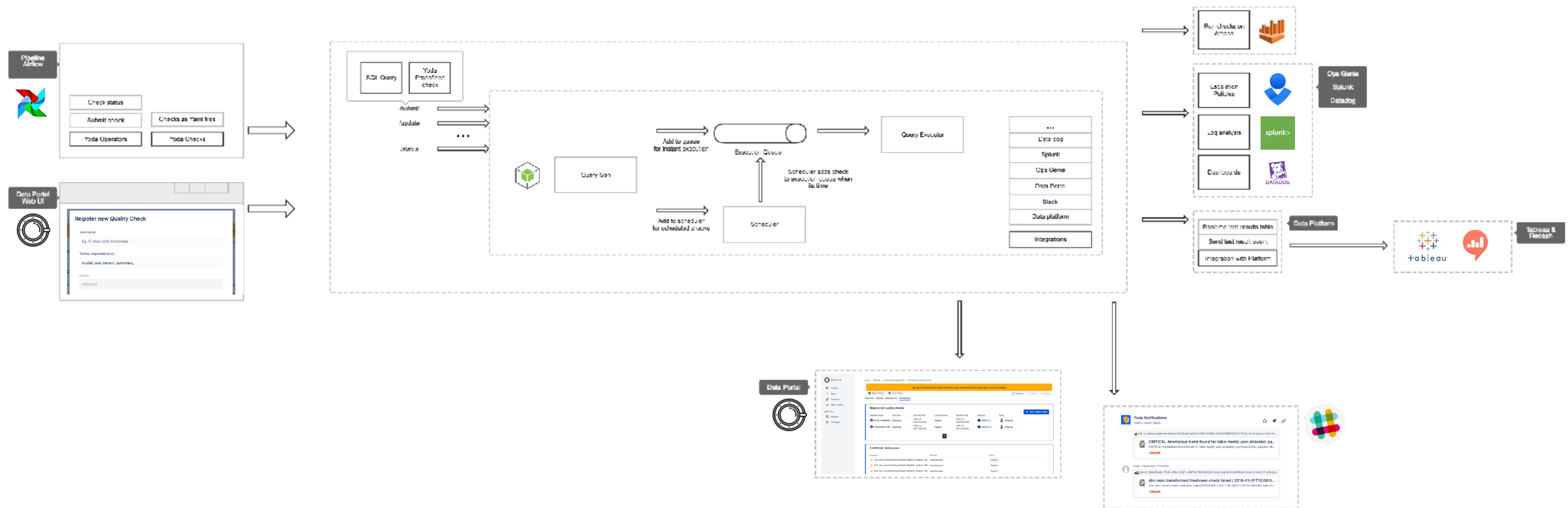




# Last bit : Yoda Core



# Bringing it together



\* don't stress your eyes

# Future



**Future**

---

**Maturity**

Subscription

Integrations



## **Mature the framework**

Framework enhancements and maintenance



## **More predefined checks**

Enhance the Yoda code gen to support more predefined checks. Reduce the on boarding curve.



## **Nurture DQ culture**

Nurture the Data Quality culture. Know that the data can be bad, and be aware when it is.

**Future**

---

Maturity

**Subscription**

Integrations



## **Subscription model**

Ability to subscribe to existing checks and not having to create new checks. Support multiple stake holders.

**Future**

---

Maturity

Subscription

**Integrations**

## **MOAR Integrations !**

Better integrations with Data Portal, and other internal + external services

## **Ops Genie**

Full integration with Ops Genie

## **Clearer data lineage**

And propagating data issues to reports and table lineage



# Thank you!



YASH SHARMA | SENIOR DATA ENGINEER | ATLASSIAN