dojo

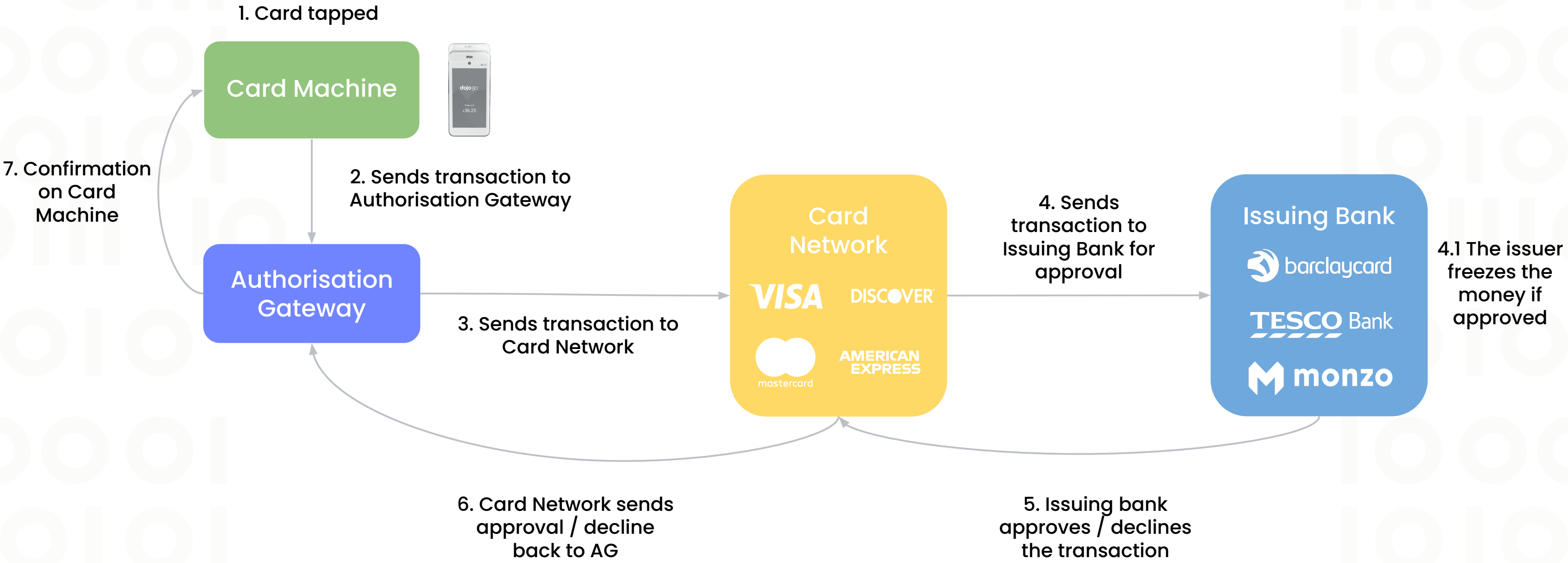
Building a business-critical data platform to process over £50bn in card transactions

www.dojo.tech

One of the largest Fintechs in Europe

Enabling 130,000 businesses to take card payments from 4 million consumers per day

Taking a card payment is complicated and highly regulated



1. Card tapped

Card Machine

2. Sends transaction to Authorisation Gateway

Authorisation Gateway

3. Sends transaction to Card Network

Card Network

4. Sends transaction to Issuing Bank for approval

Issuing Bank

4.1 The issuer freezes the money if approved

5. Issuing bank approves / declines the transaction

6. Card Network sends approval / decline back to AG

7. Confirmation on Card Machine

Challenges

Building a nuclear power station - it cannot fail

Regulatory

FCA FINANCIAL CONDUCT AUTHORITY

Safeguard customer funds from working capital at all times

PCI Security Standards Council

Data contains raw card numbers (PANs)

Complexity

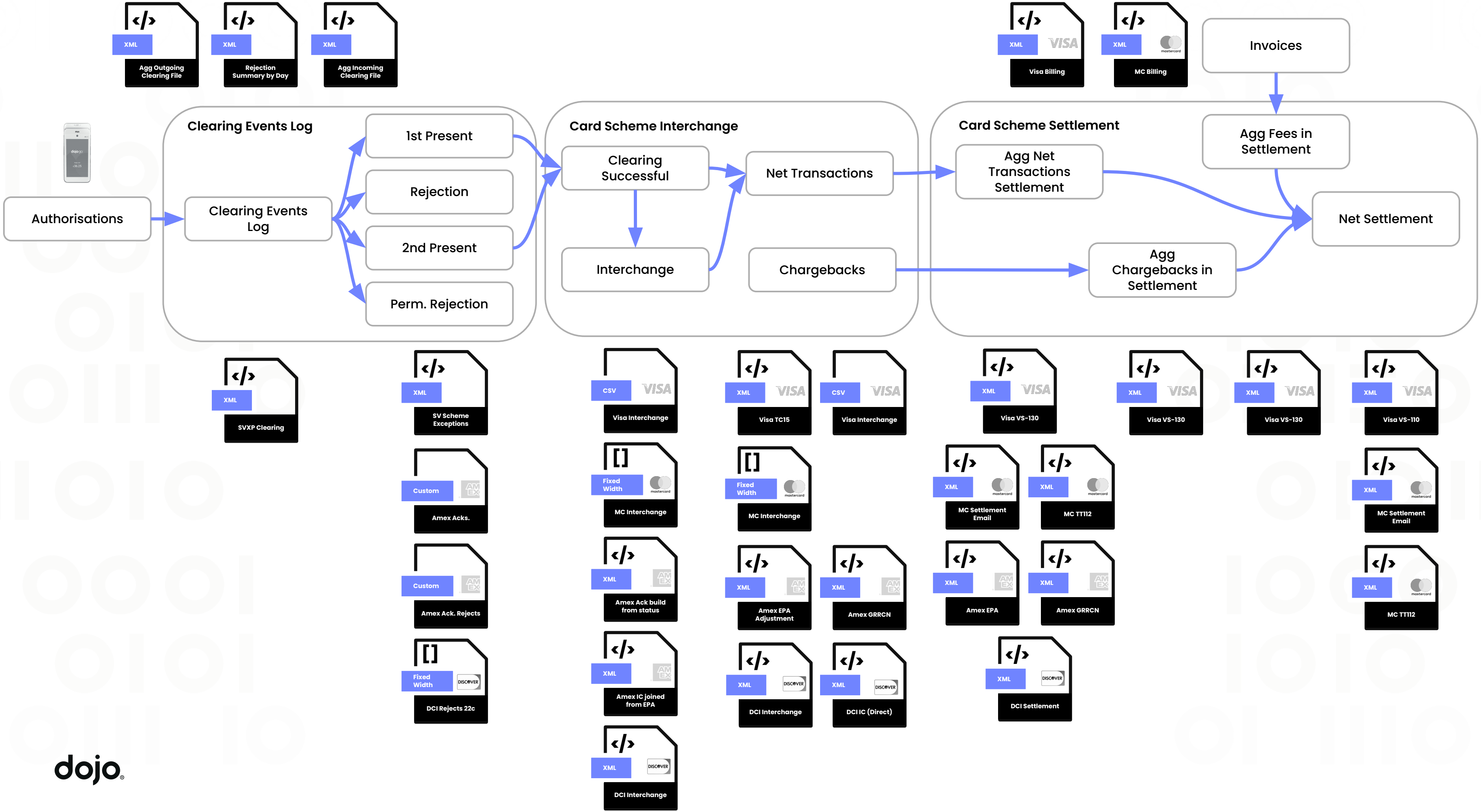
Over 1000+ schemas
Vast number of different proprietary file formats that are fragile and frequently change

Multiple file formats and sizes
Files vary from few megabyte XML all the way through to multi-gigabyte proprietary files with tens of millions of rows.

Scalability

Unpredictable demand
Unpredictable surges in transaction volumes

Business Critical (99.99%)-
Internally we refer to this as building a nuclear power station - it cannot fail.



XML Agg Outgoing Clearing File

XML Rejection Summary by Day

XML Agg Incoming Clearing File

XML Visa Billing

XML MC Billing

Invoices

Authorisations

Clearing Events Log

Clearing Events Log

1st Present

Rejection

2nd Present

Perm. Rejection

Card Scheme Interchange

Clearing Successful

Interchange

Net Transactions

Chargebacks

Card Scheme Settlement

Agg Net Transactions Settlement

Agg Fees in Settlement

Net Settlement

Agg Chargebacks in Settlement

XML SVXP Clearing

XML SV Scheme Exceptions

Custom Amex Acks.

Custom Amex Ack. Rejects

Fixed Width DCI Rejects 22c

CSV Visa Interchange

Fixed Width MC Interchange

XML Amex Ack build from status

XML Amex IC joined from EPA

XML DCI Interchange

XML Visa TC15

CSV Visa Interchange

Fixed Width MC Interchange

XML Amex EPA Adjustment

XML Amex GRRCN

XML DCI Interchange

XML DCI IC (Direct)

XML Visa VS-130

XML MC Settlement Email

XML MC TT112

XML Amex EPA

XML Amex GRRCN

XML DCI Settlement

XML Visa VS-130

XML Visa VS-130

XML Visa VS-110

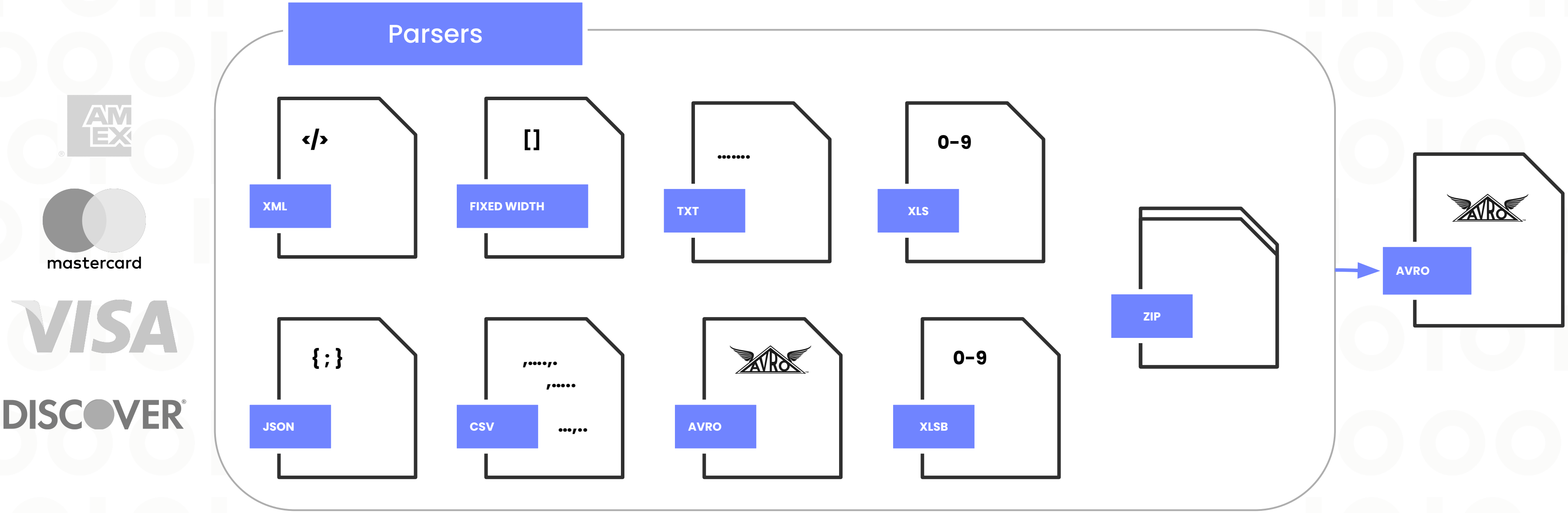
XML MC Settlement Email

XML MC TT112

dojo

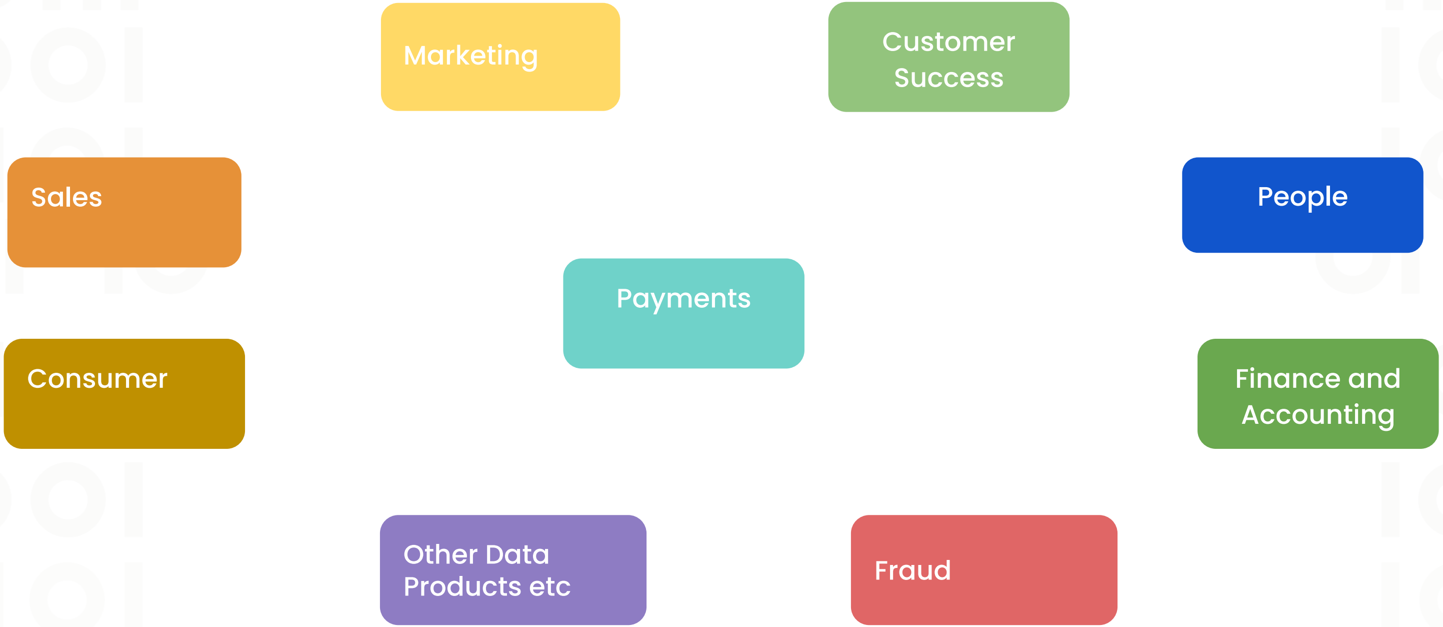
Abstracting away the complexity

Transformation into a consistent file format - Avro



dojo

other Data Domains as well ...



Marketing

Customer Success

Sales

People

Payments

Consumer

Finance and Accounting

Other Data Products etc

Fraud

Data Infrastructure Generations

Enterprise Data Warehouse

Big Data Ecosystem

Centralised Data Platform

Batch and Real Time Streaming

Cloud Based Managed Services

Challenges: Central Data Platform

Centralised Team Ownership

Siloed and Specialised Data Platform Team

Stretched Data Platform Team

Data Quality, Accountability and Democratisation

Data Quality Issues

Lack of Accountability when it comes to data issues

Delays in Data Access and Insights

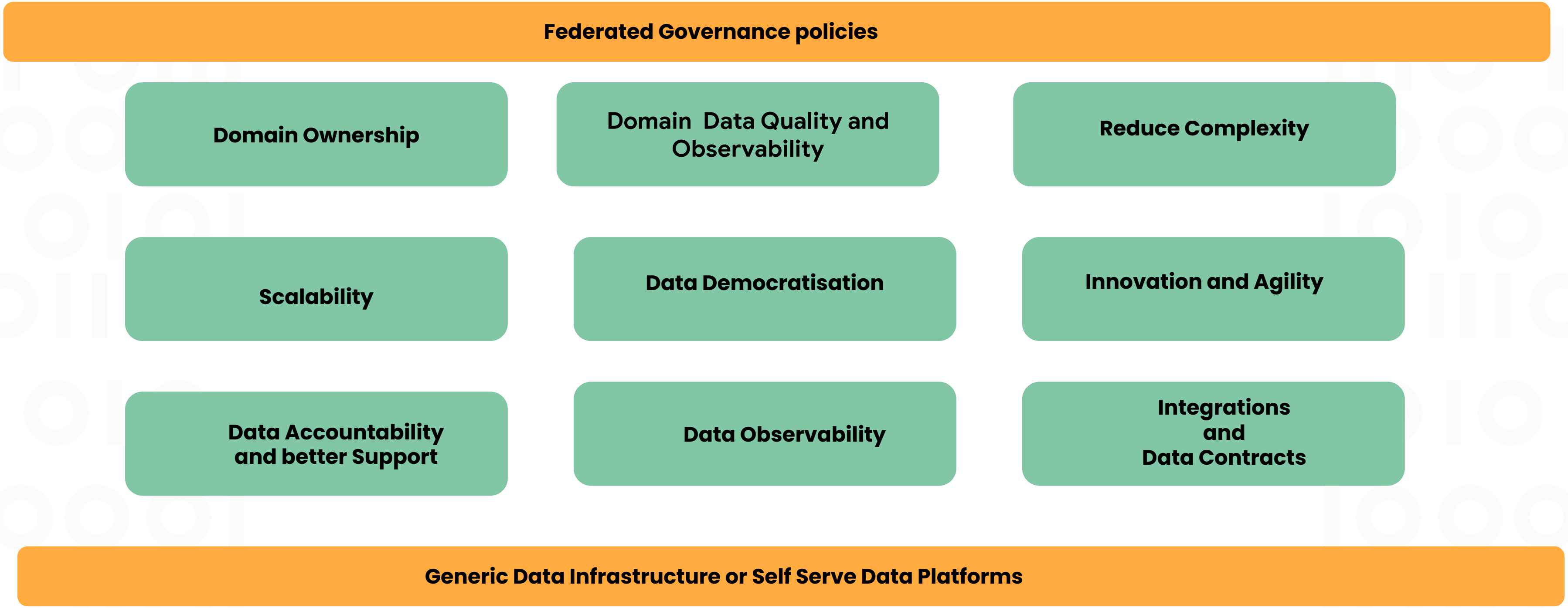
Scalability

Data Volume and New Data Sources

Cost Efficiency

dojo

What’s Next



Federated Governance policies

Domain Ownership

Domain Data Quality and Observability

Reduce Complexity

Scalability

Data Democratisation

Innovation and Agility

Data Accountability and better Support

Data Observability

Integrations and Data Contracts

Generic Data Infrastructure or Self Serve Data Platforms

Data Mesh

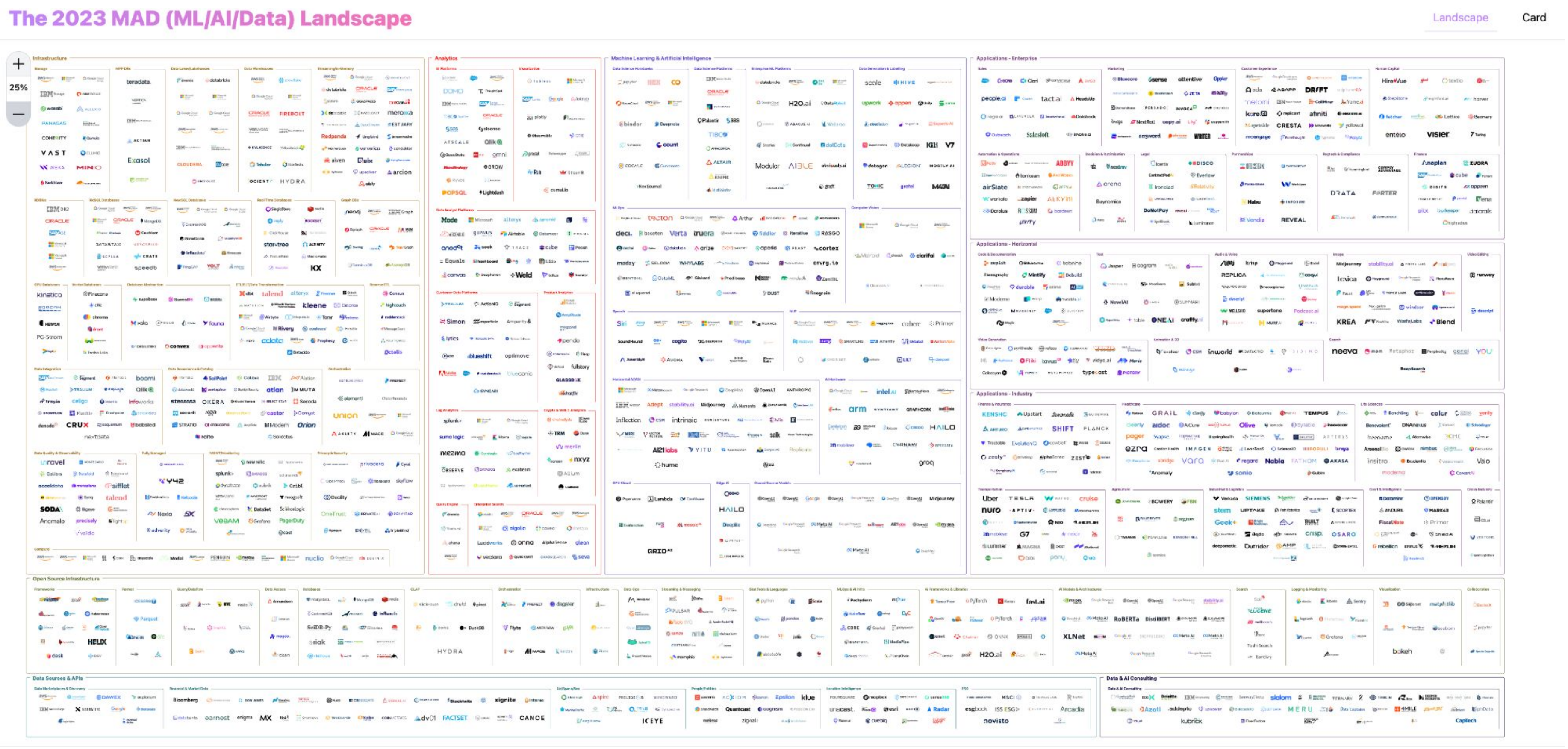
Domain Ownership

Data as a Product

Self-serve Data Platform

Federated Computational Governance

Modern Data Stack is Broken



The 2023 MAD (ML/AI/Data) Landscape

Landscape
Card
+
25%
–
Infrastructure
Analytics
Machine Learning & Artificial Intelligence
Applications - Enterprise
Applications - Horizontal
Applications - Industry
Open Source Infrastructure
Data Sources & APIs
Data & AI Consulting

What should we do ?

Build Small and Go Big

Open Source Tools

Early Feedbacks

Kubernetes is your friend

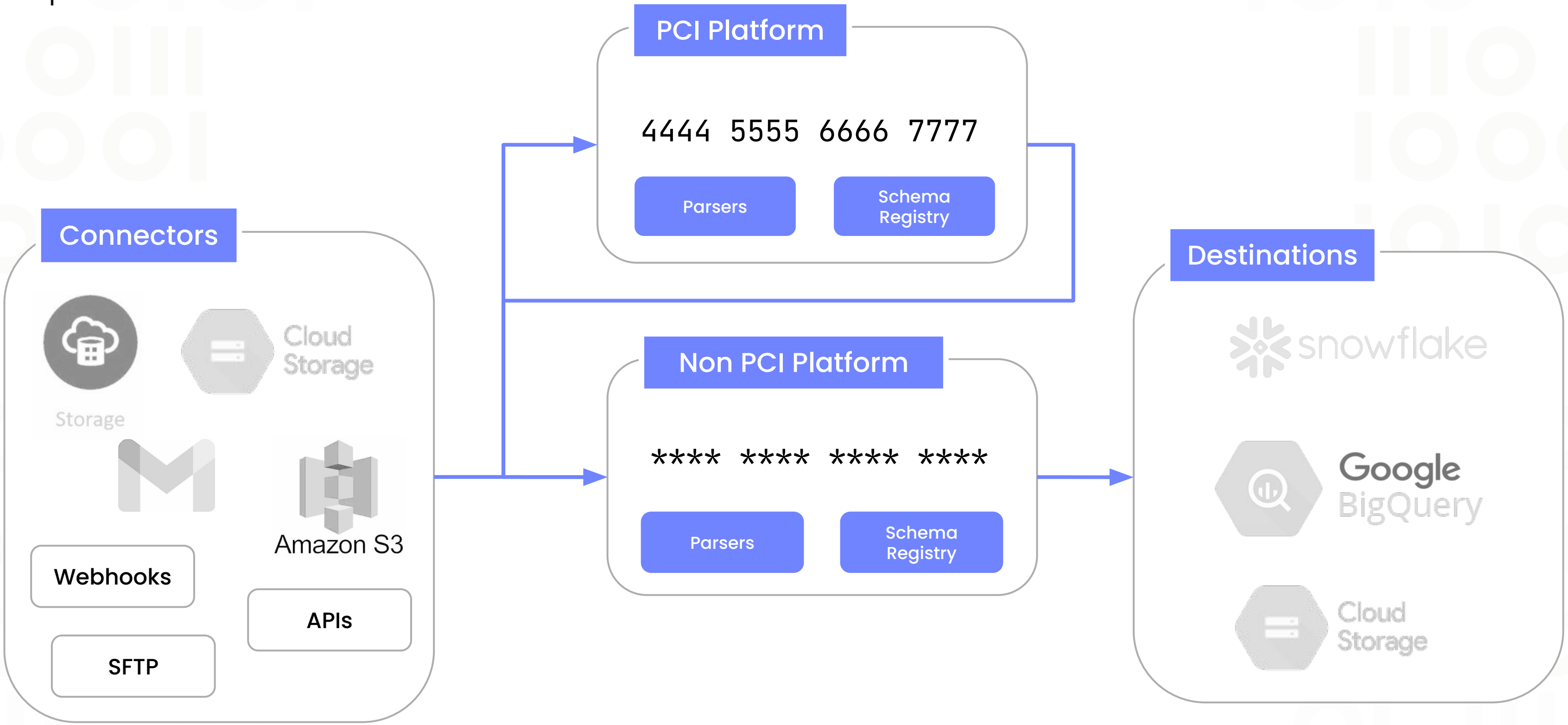
Managed Cloud Services can be handy

Cloud Agnostic

dojo

Platform Overview

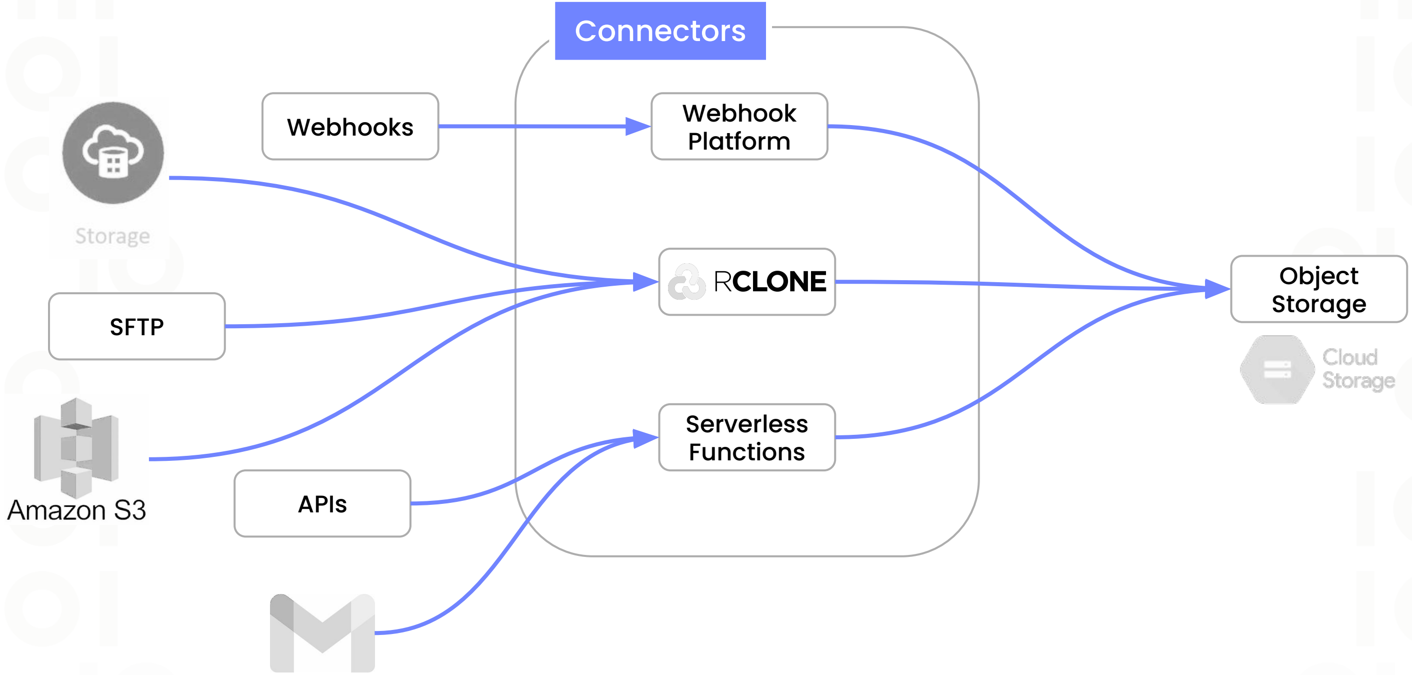
Component based architecture



dojo.

Connectors Overview

A unified approach to ingesting data from multiple sources



dojo.

PCI DSS Level 1

Process, store, or transmit credit card or cardholder data maintain a secure environment

Card data transmitted securely into the Data Platform

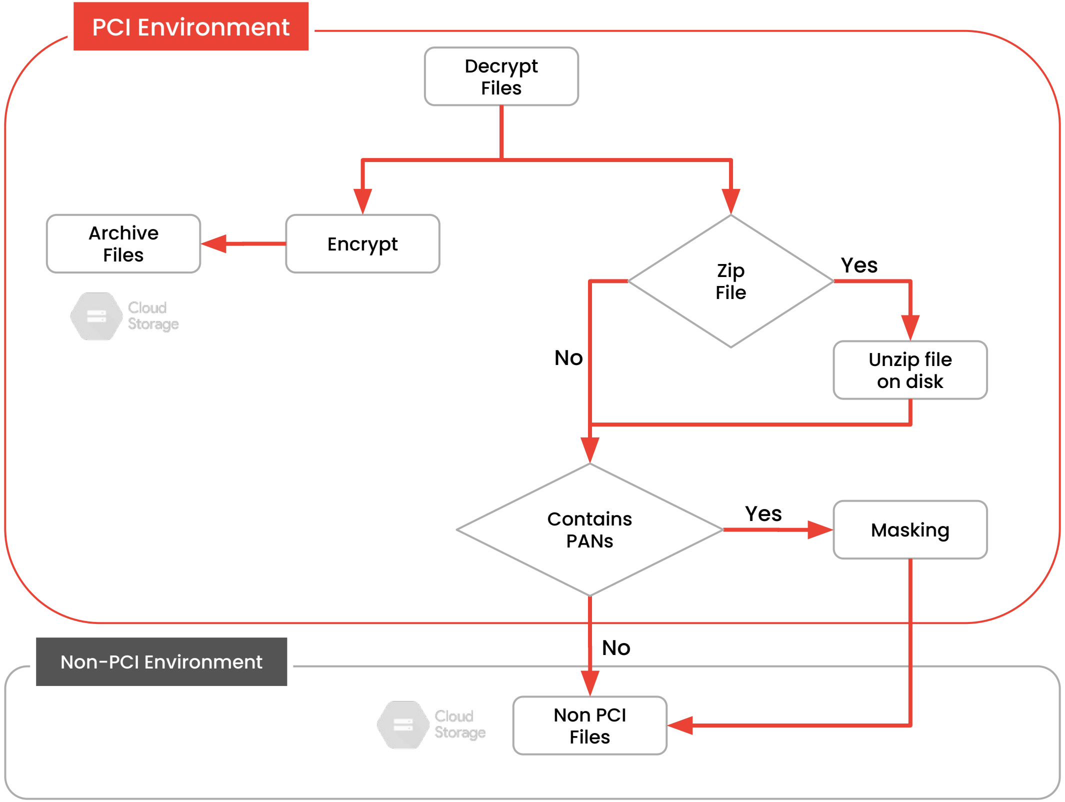
Strong Encryption, Monitoring and Security testing of Data Platform

Yearly Audits to verify the security of the Data Platform

PCi Security Standards Council®

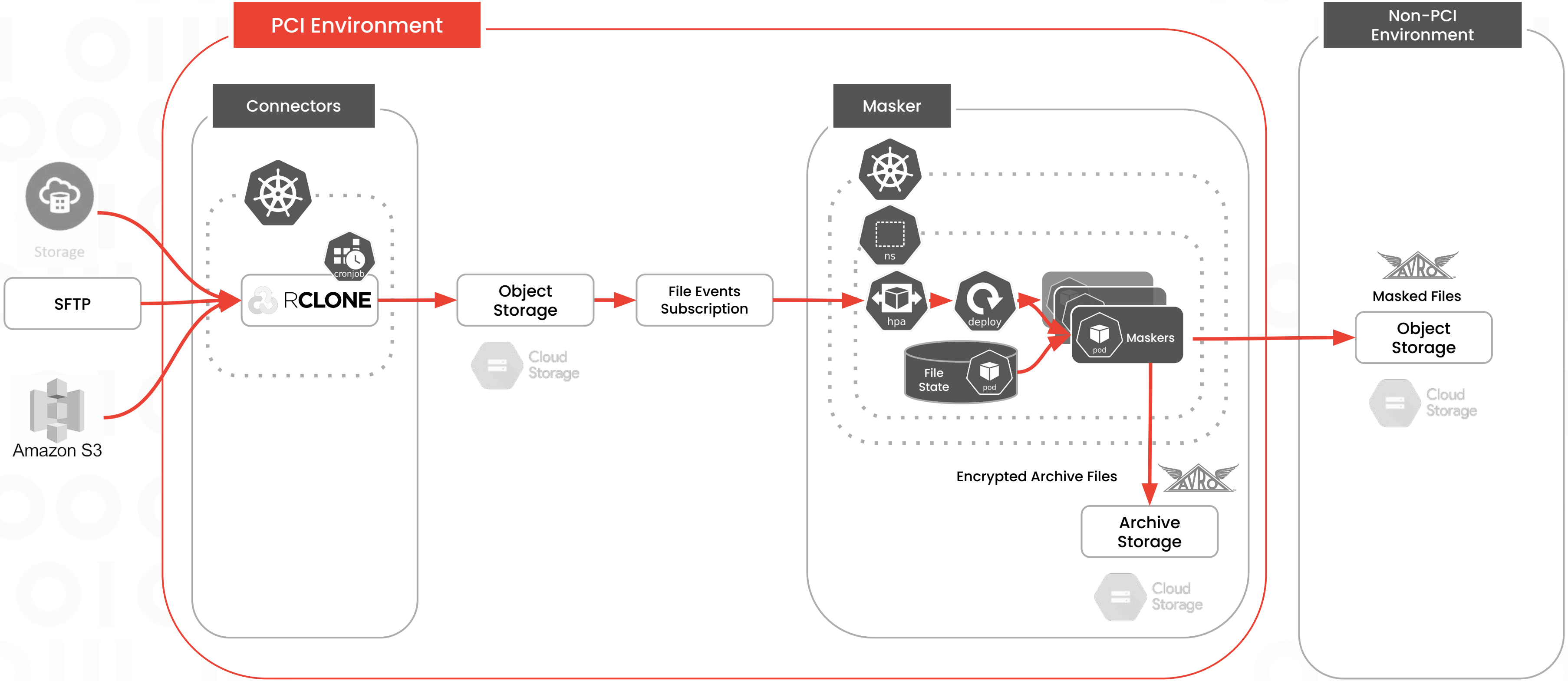
PCI Management Process

Secure and prevent the egress of PCI sensitive data



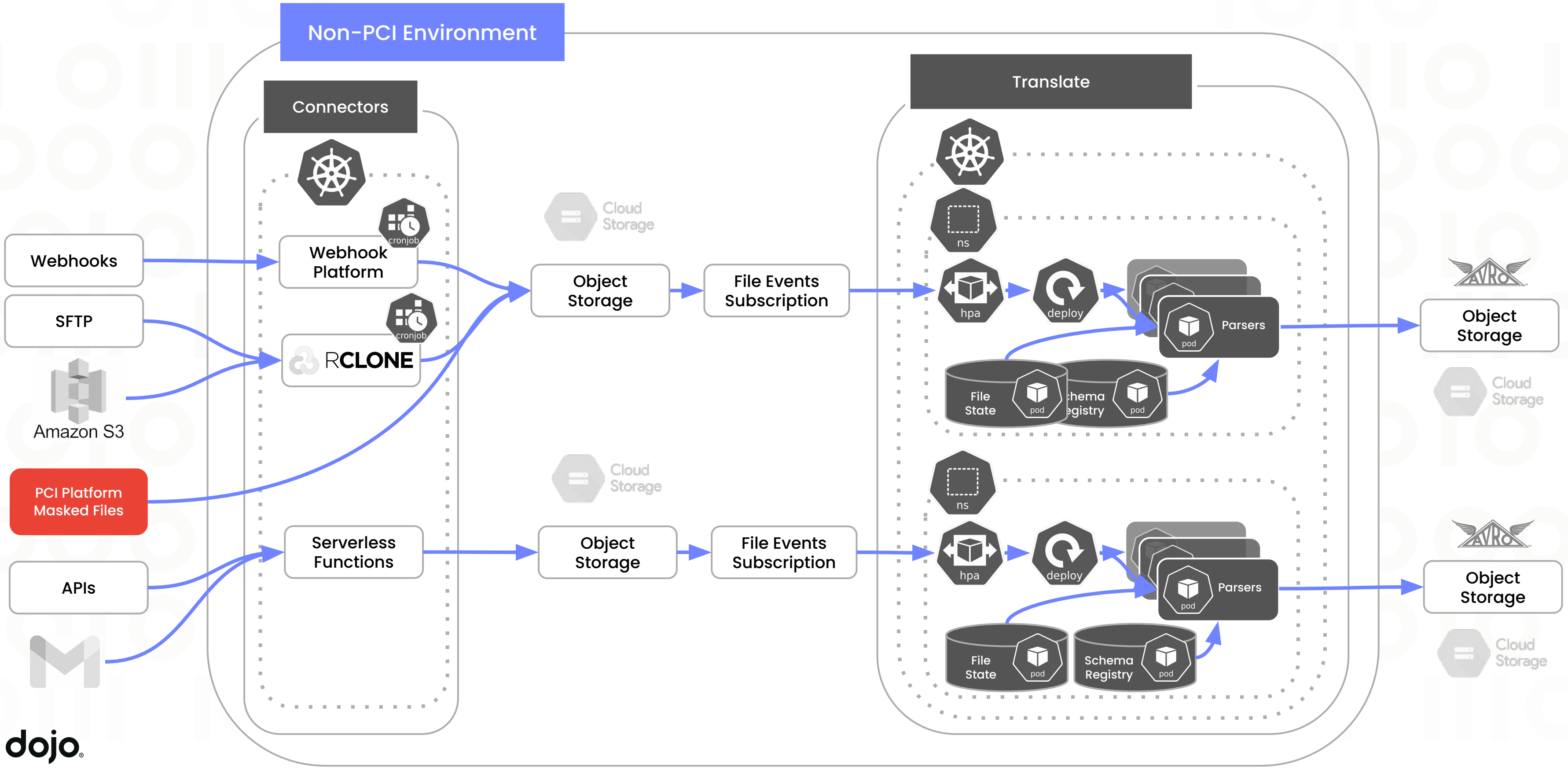
dojo.

PCI Platform Overview



dojo.

Transformation Platform Overview



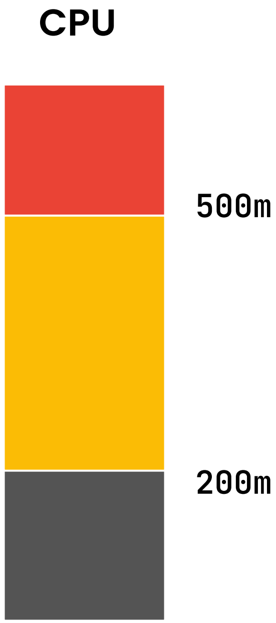
dojo.

Autoscaling Challenges

Capacity planning is essential to scaling efficiently

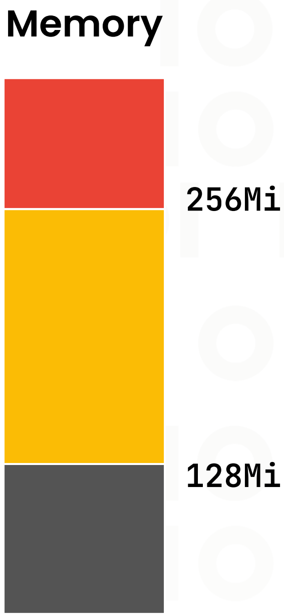
Limit
where CPU
throttling begins

Request
amount guaranteed
by k8s



Limit
where **processes are killed**

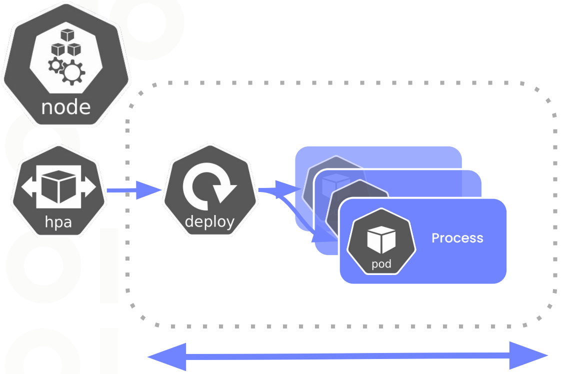
Request
amount guaranteed
by k8s



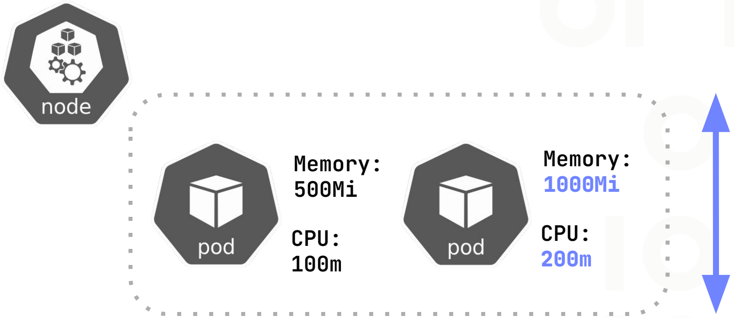
Autoscaling

Two approaches to autoscaling in Kubernetes - Horizontal (HPA) or Vertical (VPA)

Scale Out: Increase / decrease number of pods based on metrics



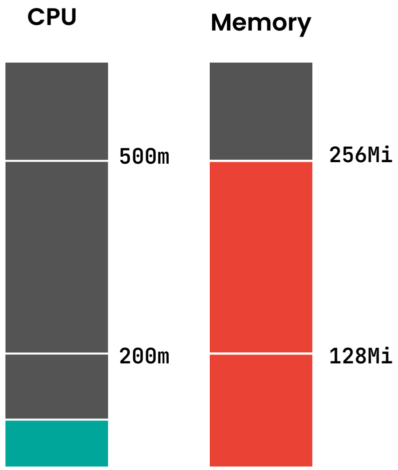
Scale Up: Increase resources assigned to workload



Autoscaling Triggers

Kubernetes provides three solutions depending on the use case

Resource Usage



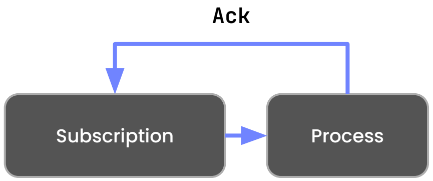
metrics.k8s.io

Custom Metrics

hits-per-second

custom.metrics.k8s.io

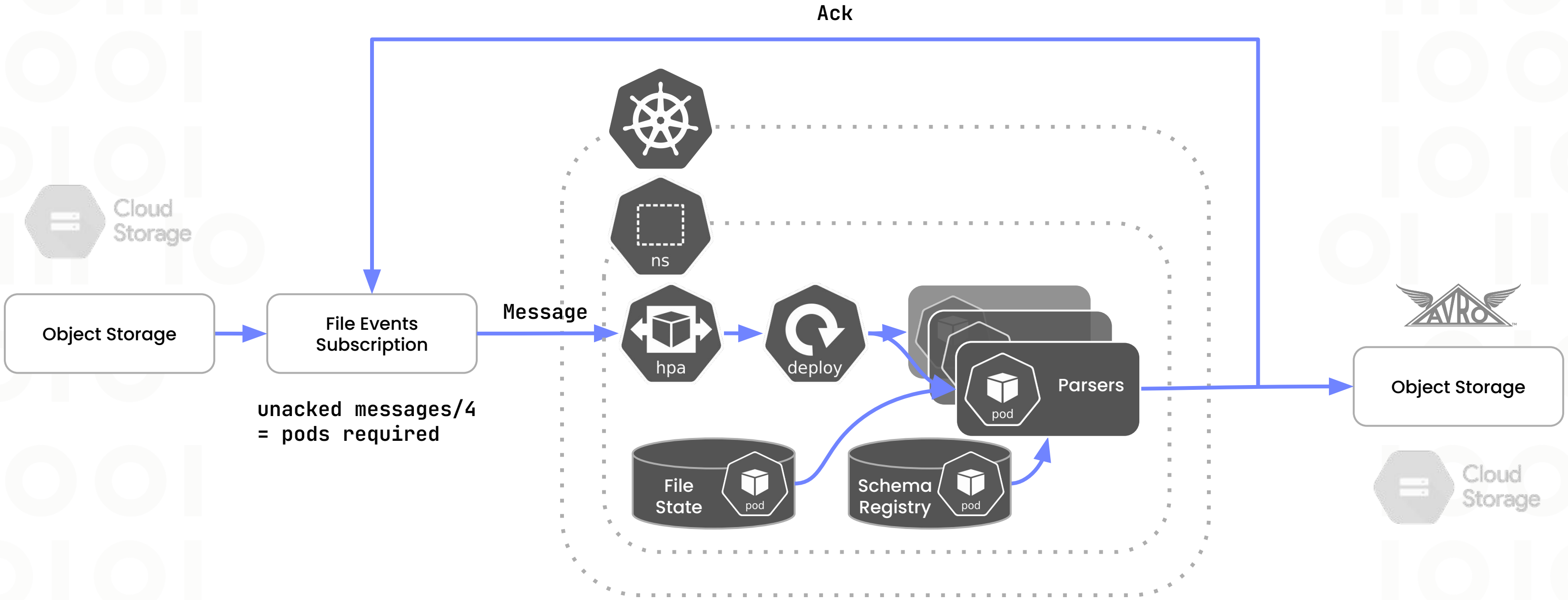
External Custom Metrics



external.metrics.k8s.io

Using HPA to elastically scale

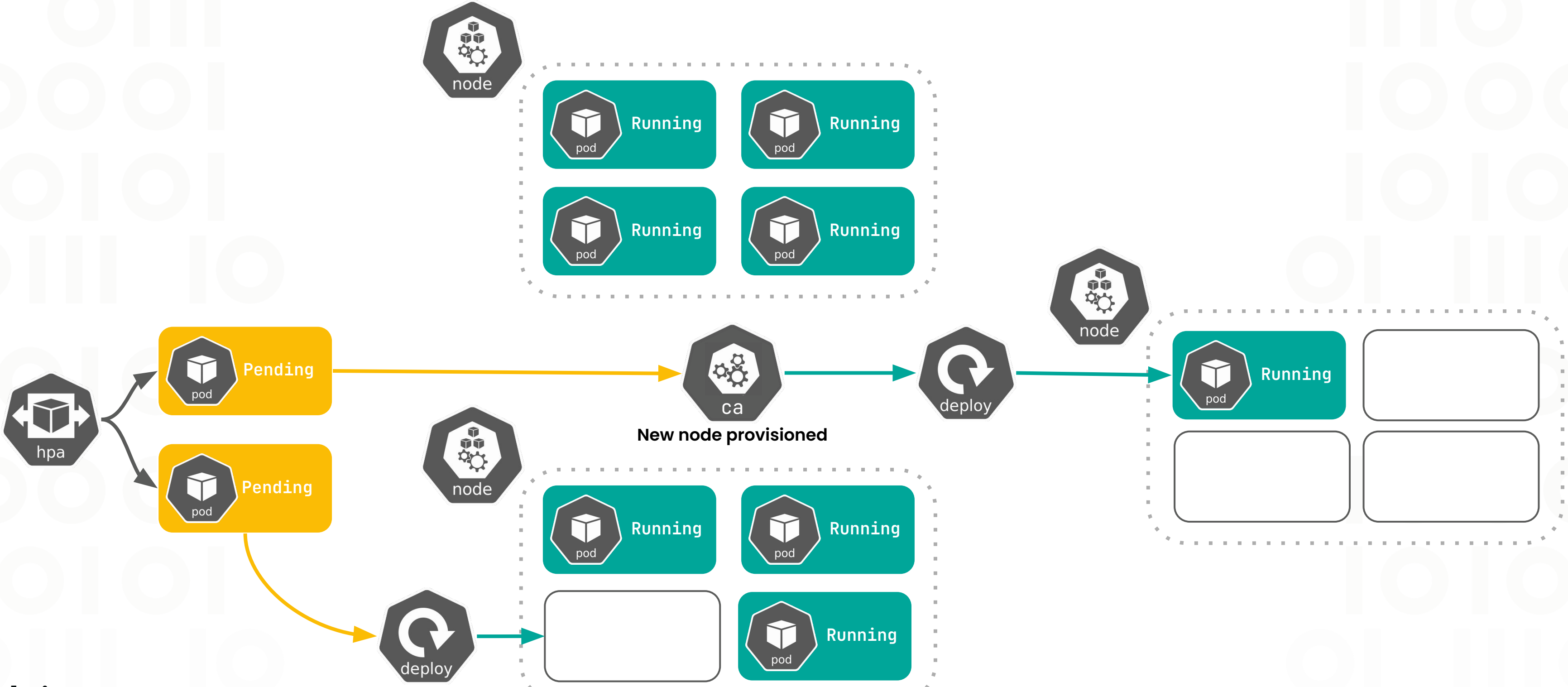
Pods required determined by number of unacknowledged messages in queue



dojo.

Cluster Autoscaler

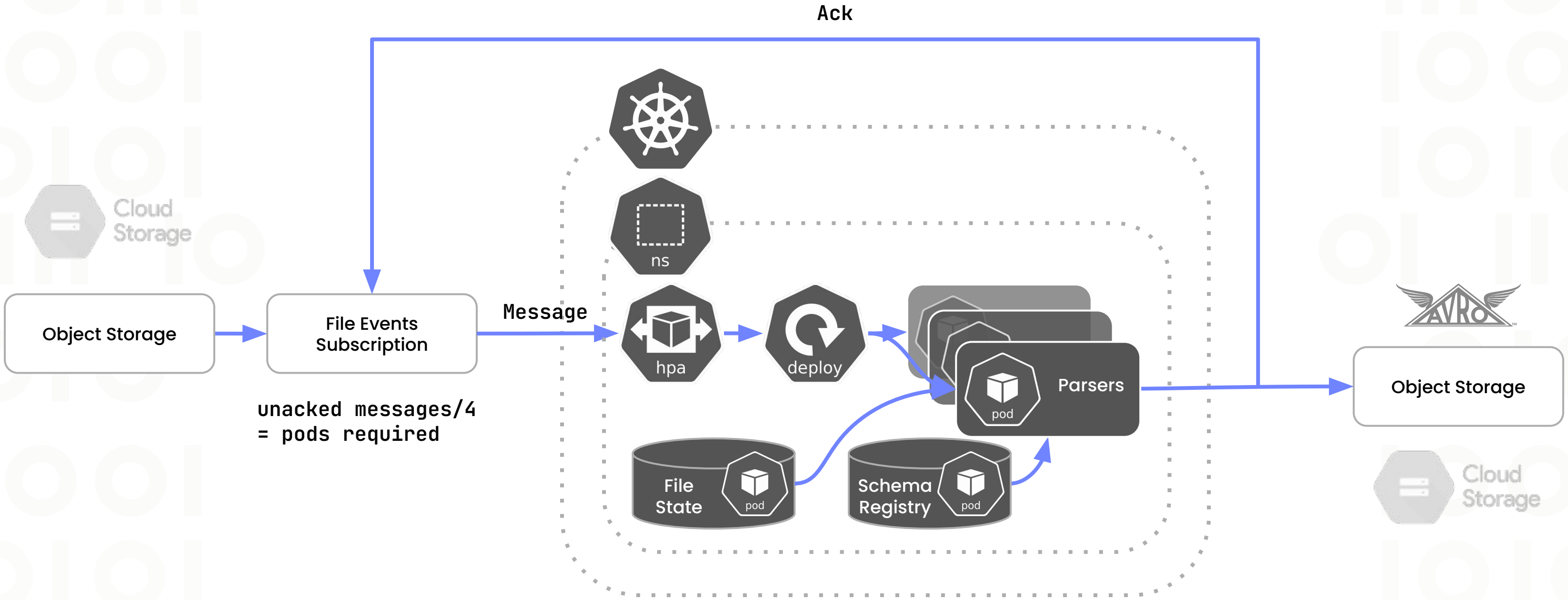
New nodes provisioned based on pods in **`Pending`** state



dojo

Using HPA to elastically scale

Pods required determined by number of unacknowledged messages in queue

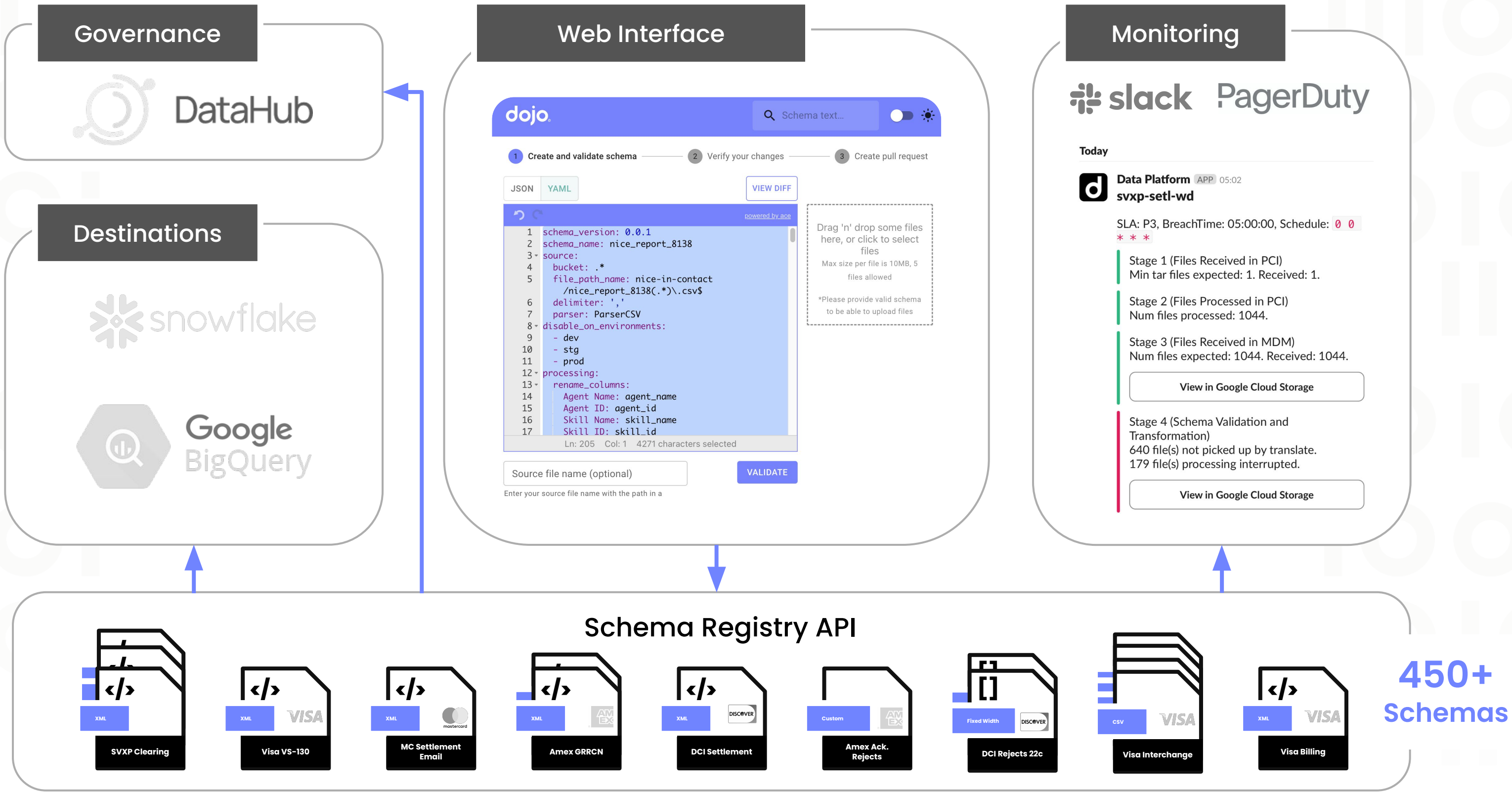


unacked messages/4
= pods required

dojo.

Schema Registry

A central source of metadata for file lifecycle management



1 Create and validate schema — 2 Verify your changes — 3 Create pull request

```yaml
schema_version: 0.0.1
schema_name: nice_report_8138
source:
  bucket: .*
  file_path_name: nice-in-contact
    /nice_report_8138(.*)\.csv$
  delimiter: ','
  parser: ParserCSV
disable_on_environments:
  - dev
  - stg
  - prod
processing:
  rename_columns:
    Agent Name: agent_name
    Agent ID: agent_id
    Skill Name: skill_name
    Skill ID: skill_id
```

Ln: 205 Col: 1 4271 characters selected

Drag 'n' drop some files here, or click to select files

Max size per file is 10MB, 5 files allowed

*Please provide valid schema to be able to upload files

Enter your source file name with the path in a

Today

Data Platform APP 05:02

svxp-setl-wd

SLA: P3, BreachTime: 05:00:00, Schedule: `0 0 * * *`

Stage 1 (Files Received in PCI)
Min tar files expected: 1. Received: 1.

Stage 2 (Files Processed in PCI)
Num files processed: 1044.

Stage 3 (Files Received in MDM)
Num files expected: 1044. Received: 1044.

View in Google Cloud Storage

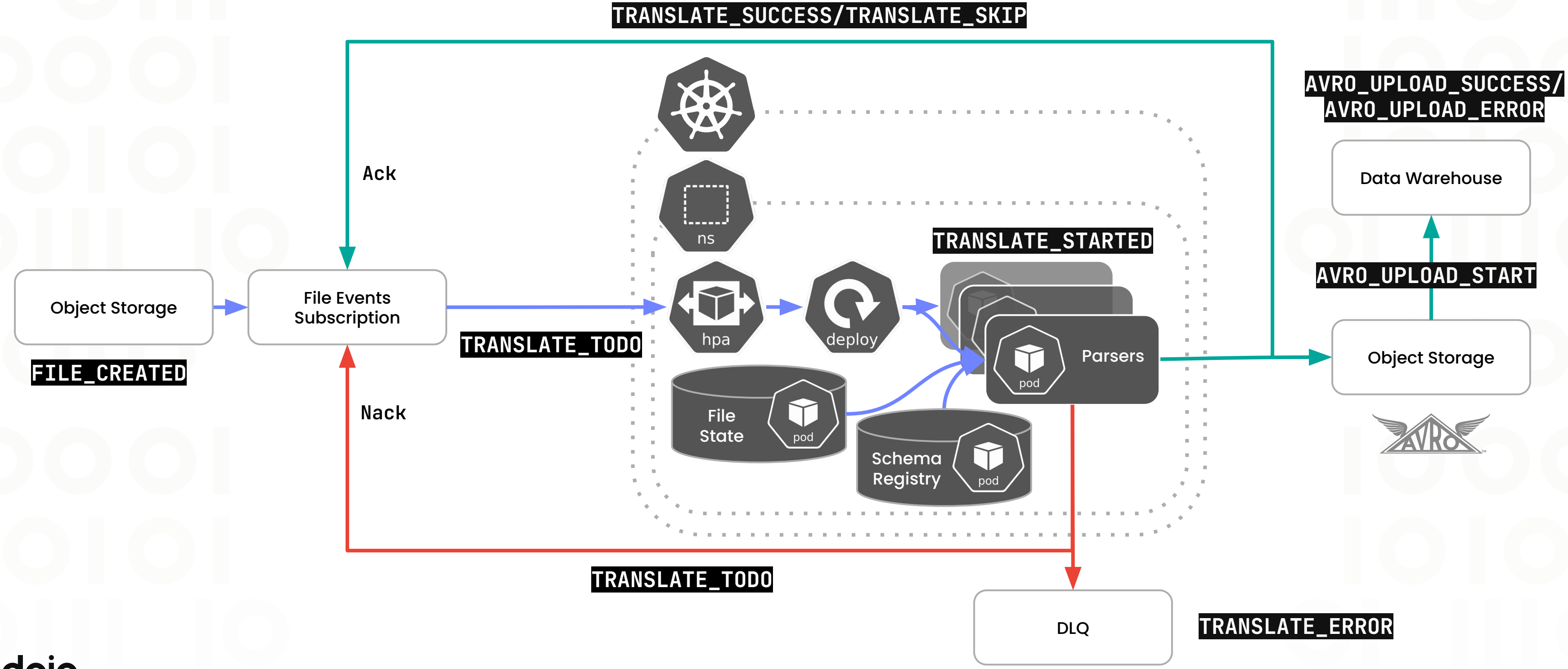
Stage 4 (Schema Validation and Transformation)
640 file(s) not picked up by translate.
179 file(s) processing interrupted.

View in Google Cloud Storage

dojo.

State Management

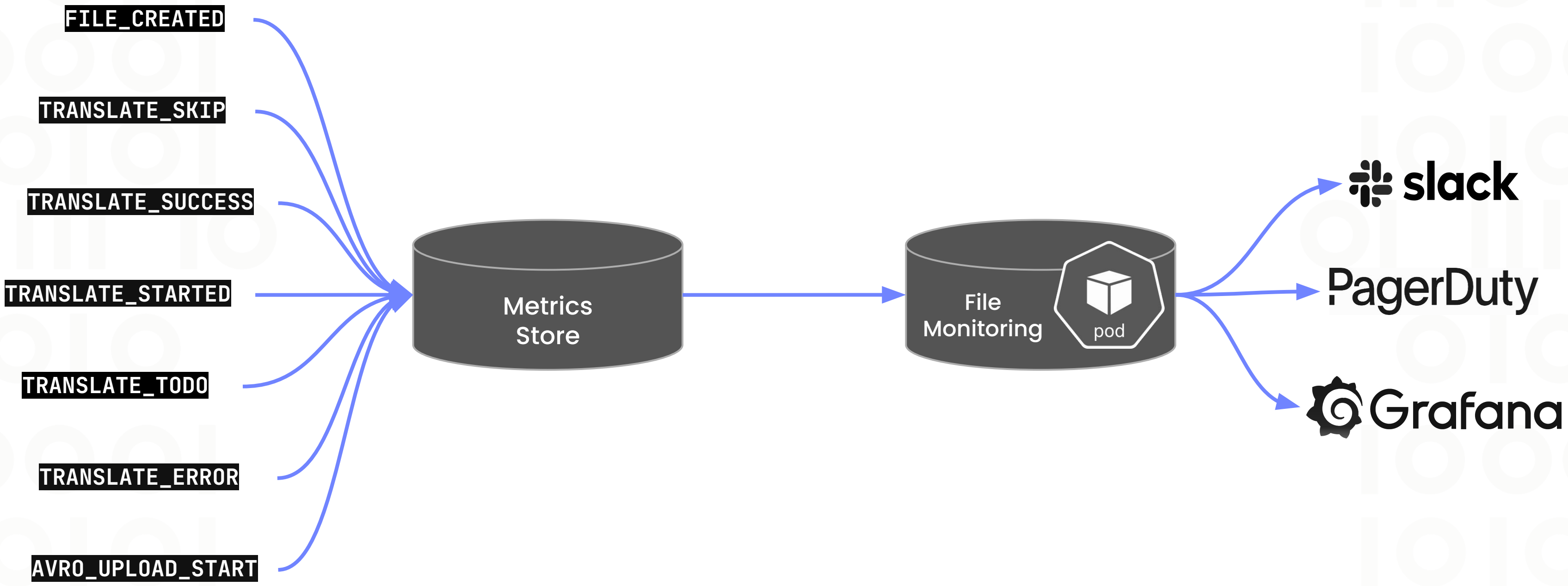
State of every file recorded throughout the transformation process



dojo.

Real time File Monitoring

State of every file recorded throughout the transformation process



dojo

E2E File Monitoring

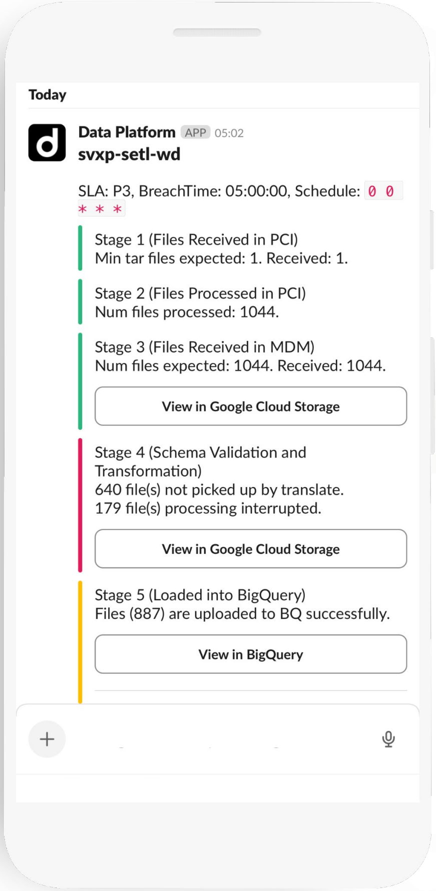
PagerDuty

P0 P1

slack

P2 P3

Looker



Today

Data Platform APP 05:02

svxp-setl-wd

SLA: P3, BreachTime: 05:00:00, Schedule: `0 0 * * *`

Stage 1 (Files Received in PCI)
Min tar files expected: 1. Received: 1.

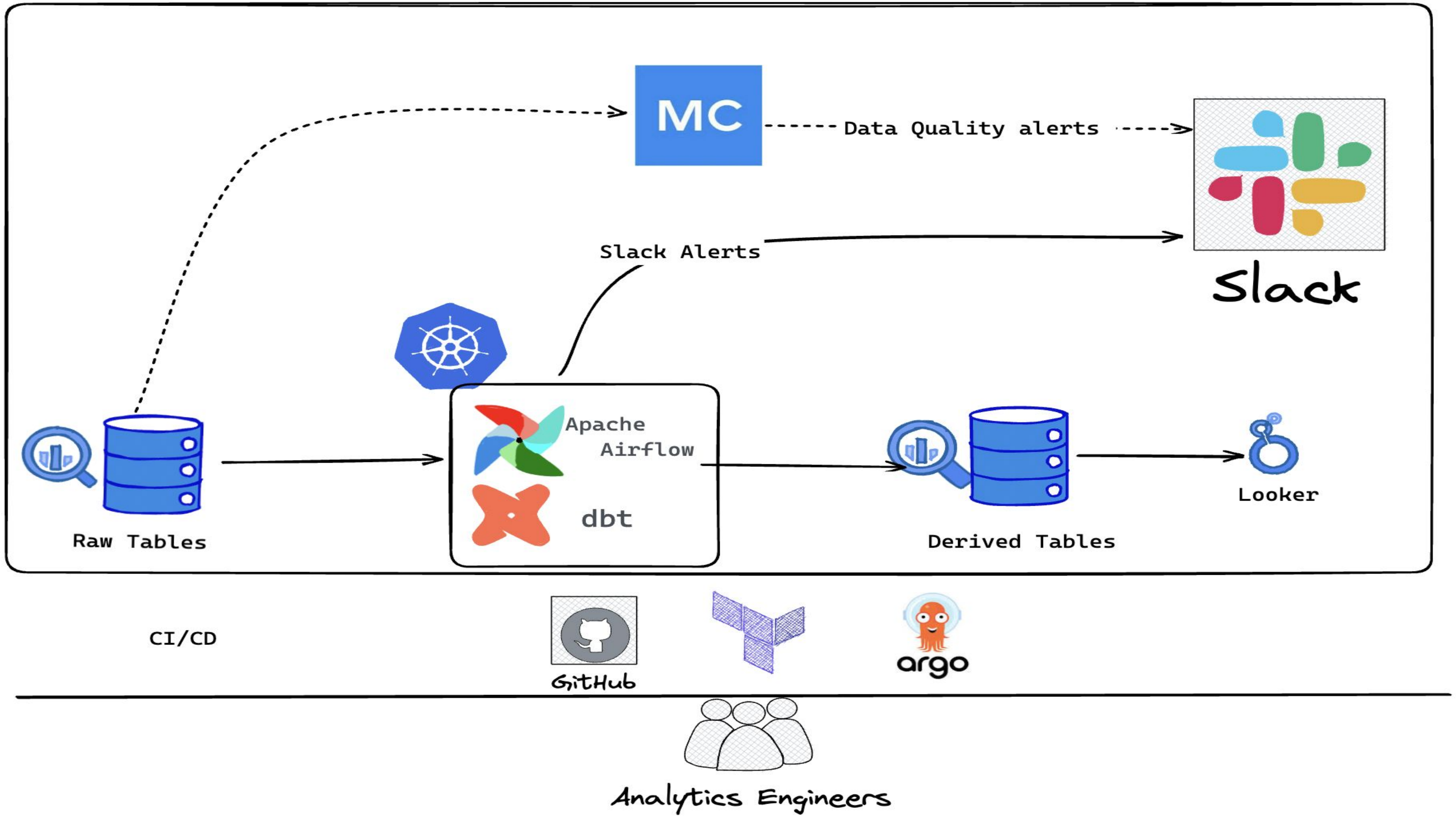
Stage 2 (Files Processed in PCI)
Num files processed: 1044.

Stage 3 (Files Received in MDM)
Num files expected: 1044. Received: 1044.

Stage 4 (Schema Validation and Transformation)
640 file(s) not picked up by translate.
179 file(s) processing interrupted.

Stage 5 (Loaded into BigQuery)
Files (887) are uploaded to BQ successfully.

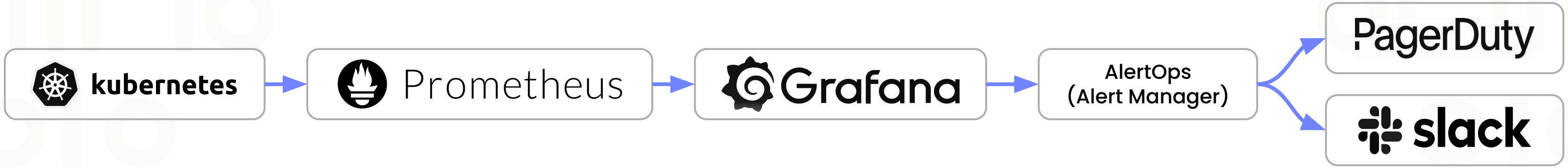
Analytics Platform



dojo

Infrastructure Observability

Monitoring and alerting on Data Platform resources



Challenges: Embracing Data Mesh

Cultural Shift

Talent and Skill Gaps

Data Governance Complexity

Data Mesh Tooling

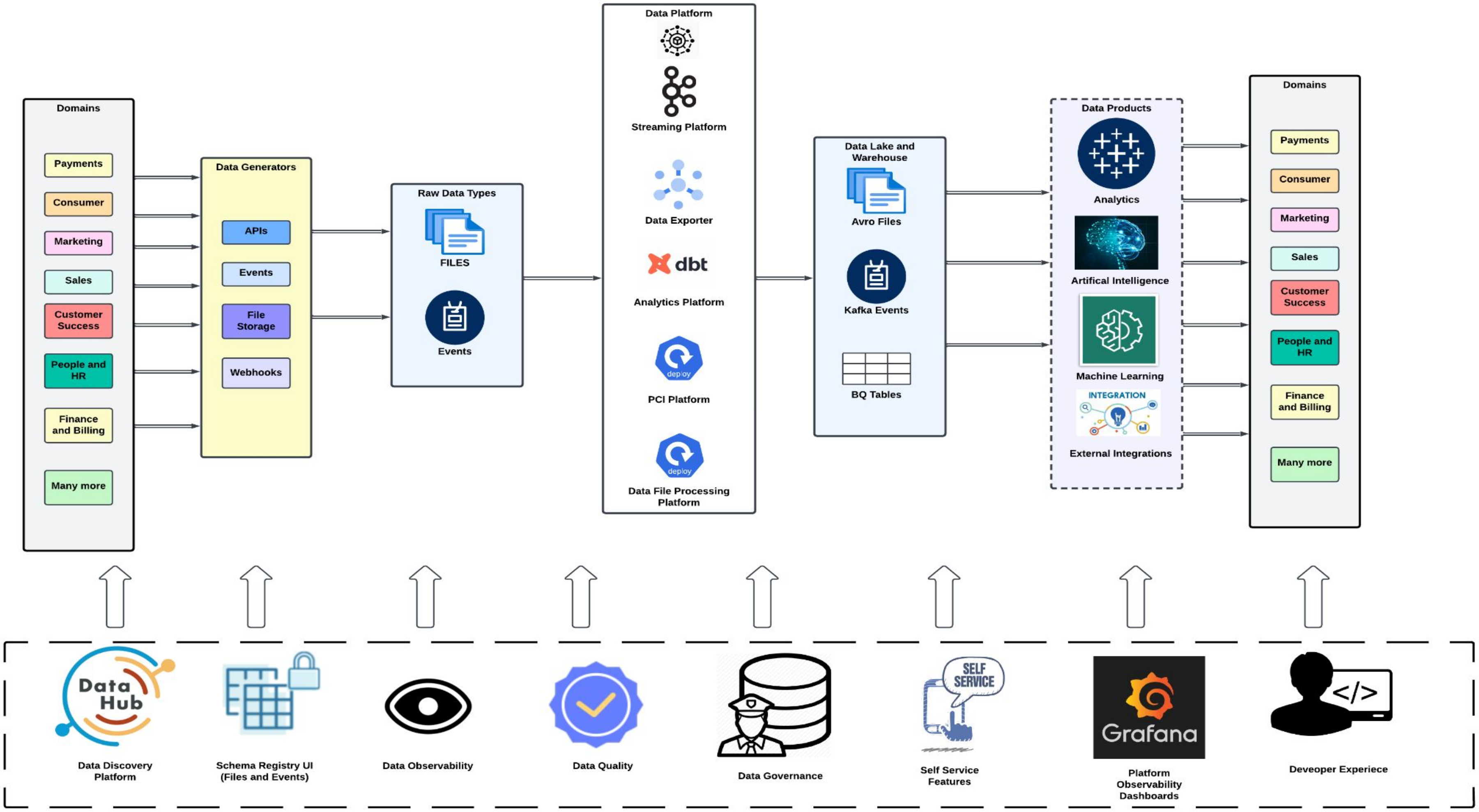
Data Privacy and Security

Data Integrations

ROI Measurement

Data Cataloging
and
Data Discovery

Legacy system Integrations


Domains
Payments
Consumer
Marketing
Sales
Customer Success
People and HR
Finance and Billing
Many more
Data Generators
APIs
Events
File Storage
Webhooks
Raw Data Types
FILES
Events
Data Platform
Streaming Platform
Data Exporter
dbt
Analytics Platform
deploy
PCI Platform
deploy
Data File Processing Platform
Data Lake and Warehouse
Avro Files
Kafka Events
BQ Tables
Data Products
Analytics
Artifical Intelligence
Machine Learning
INTEGRATION
External Integrations
Domains
Payments
Consumer
Marketing
Sales
Customer Success
People and HR
Finance and Billing
Many more
Data Hub
Data Discovery Platform
Schema Registry UI (Files and Events)
Data Observability
Data Quality
Data Governance
SELF SERVICE
Self Service Features
Grafana
Platform Observability Dashboards
</>
Deveoper Experiece

Thank you for your time!

WE'RE HIRING!

https://www.dojo.careers

dojo®

Sandeep Mehta

Data Platform Lead

in *sandeep-mehta26*