Protecting the Data Lake

A bit about Ash Narkar !

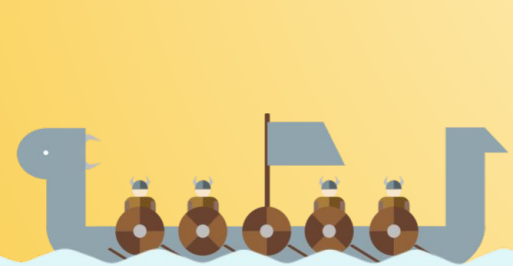
@ashtalk

styra

Open Policy Agent

Agenda

- Data Lake Overview
- Open Policy Agent
  - Community
  - Features
  - Use Cases
- Use case deep dive
  - Ceph Data Protection



Data is King !

Data is King !

- Pervasive
- Abundant
- Customer Experience
- Revenue Growth

Data is King !

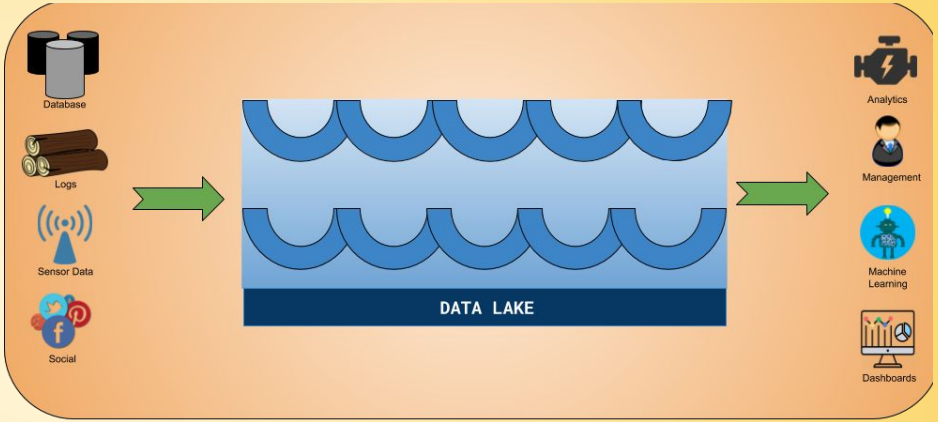
- Pervasive
- Abundant
- Customer Experience
- Revenue Growth

- Cyber Attacks
- Breaches
- Fines
- Loss of Customer Trust

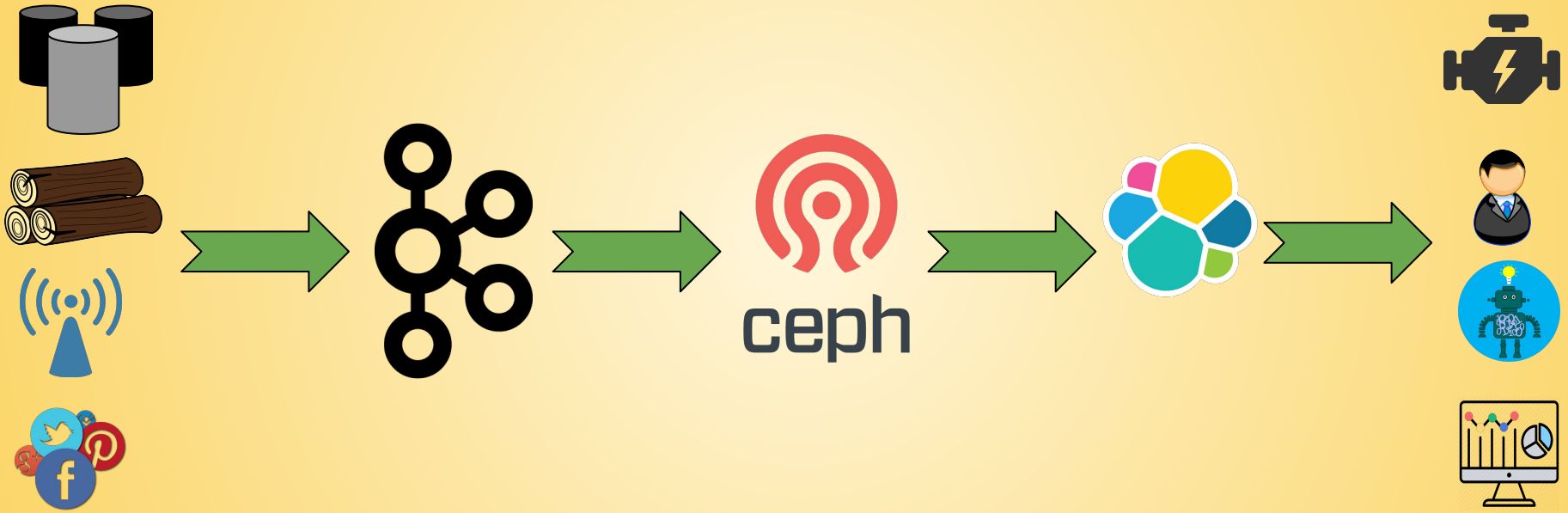
What Is A Data Lake?

Data Lake Features

- Centralized Content
- Scalability
- Multiple data type support
- Resource optimization



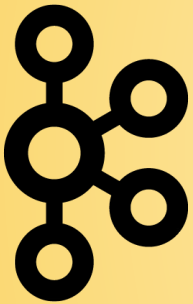
Data Lake Platform



Sources

Consumers

Data Lake Platform: Kafka



Features

- Distributed streaming platform
- Building real-time streaming data pipelines and applications

Security Challenges

- Authorization using Access Control Lists(ACLs)
- How to authorize requests based on context, like user, IP, common name in certificate

Security Policies

- Consumers of topics containing PII must be whitelisted
- Producers to topics with high fanout must be whitelisted

Data Lake Platform: Ceph

Features

- Unified distributed storage system
- Delivers object, block, and file storage

Security Challenges

- Security protocol handles only Ceph clients and servers. NO human users or applications 😔

Security Policies

- Users can access only those buckets belonging to the same geographical region as them
- Access based on a user's Business Unit, Department etc.

Data Lake Platform: Elasticsearch

Features

- Full-text search and analytics engine
- Store, search and analyze

Security Challenges

- Authorization is not considered as part of job
- User responsible for implementing access control

Security Policies

- Access control policies for a patient's PHI

Security Challenge Overview

- Distinct systems
- Changing security requirements

❌ Hardcoding policy

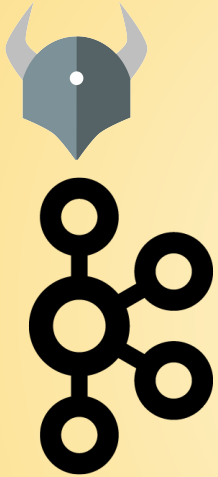
❌ Tight coupling

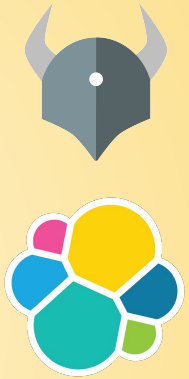
✔️ Expressiveness

✔️ Speed and performance

✔️ Unified Solution

Who can solve the Security Challenge ?





What Is OPA?

OPA: Community

Inception

Project started in 2016 at Styra.

Goal

Unify policy enforcement across the stack.

Users

Netflix
Medallia
Chef
Cloudflare
State Street
Pinterest
Intuit
Capital One
...and many more.

Use Cases

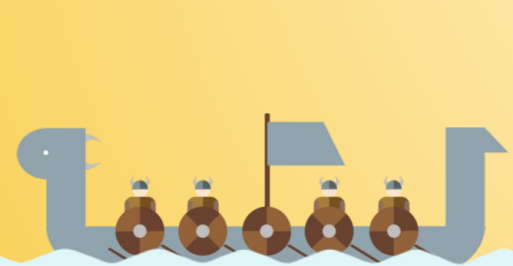
Admission control
Authorization
- ACLs
- RBAC
- IAM
- ABAC

Risk management
Data Protection
Data Filtering

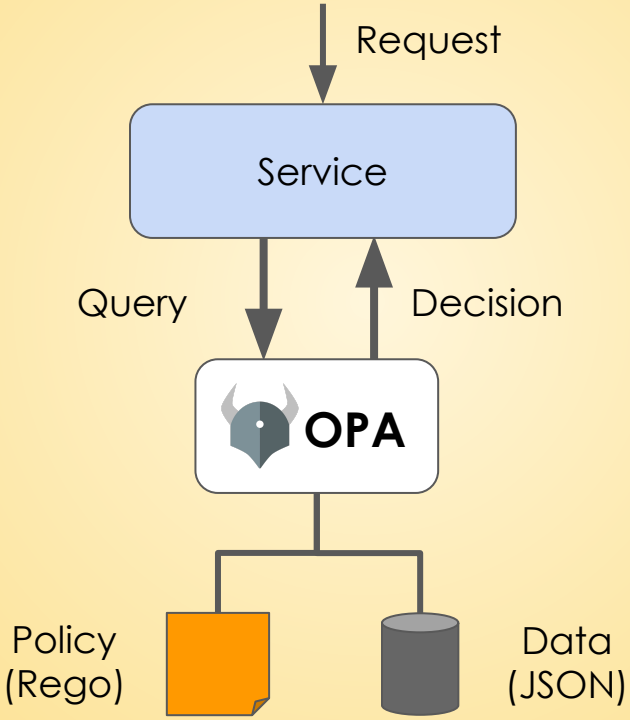
Today

CNCF project (Incubating)

59 contributors
800+ slack members
2000+ stars
20+ integrations

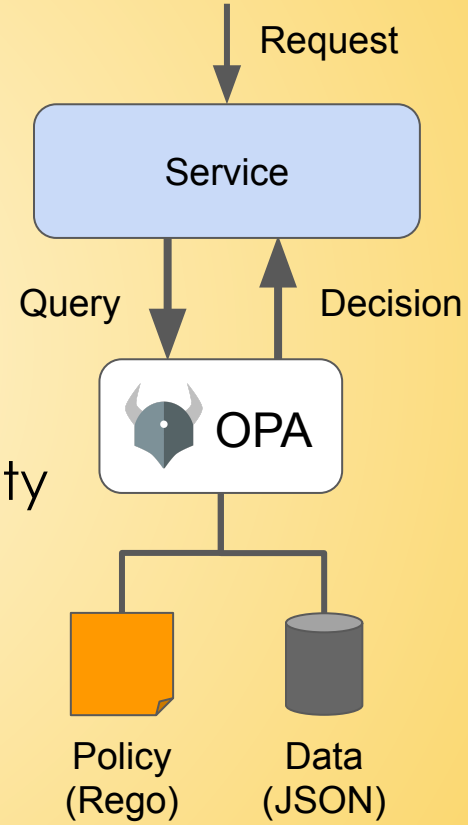


OPA: General-purpose policy engine



OPA: Features

- Declarative Policy Language (Rego)
  - Can user X do operation Y on resource Z?
  - What invariants does workload W violate?
  - Which records should bob be allowed to see?
- Library, sidecar, host-level daemon
  - Policy and data are kept in-memory
  - Zero decision-time dependencies
- Management APIs for control & observability
  - Bundle service API for sending policy & data to OPA
  - Status service API for receiving status from OPA
  - Log service API for receiving audit log from OPA
- Tooling to build, test, and debug policy
  - opa run, opa test, opa fmt, opa deps, opa check, etc.
  - VS Code plugin, Tracing, Profiling, etc.

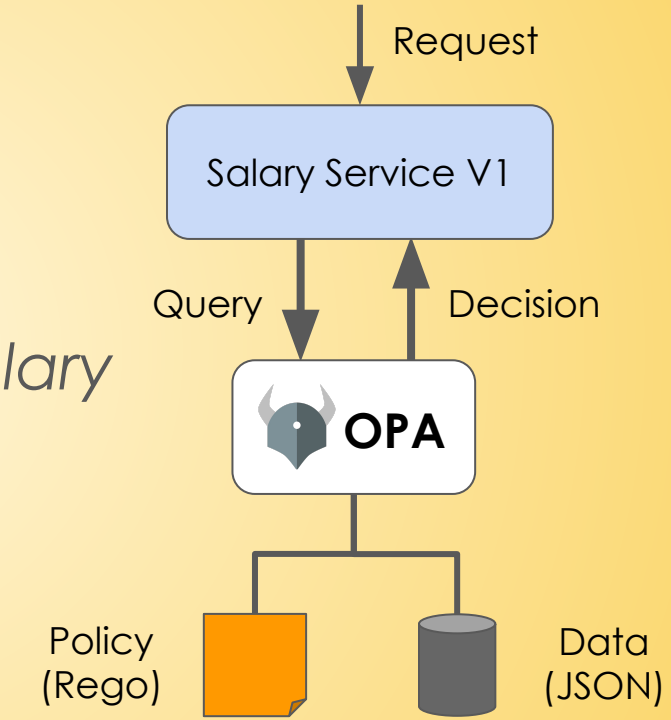


How does OPA work?

How does OPA work?

Example policy

"Employees can read their own salary and the salary of anyone they manage."



How does OPA work?

Example policy

Employees can read their own salary and the salary of anyone they manage.

Input Data

```
method: "GET"
path: ["salary", "bob"]
user: "bob"
```

3 Steps to OPA 🎉

Step 1: Clone OPA Repo

3 Steps to OPA 🎉

Step 1: Clone OPA Repo

Step 2: Build OPA binary

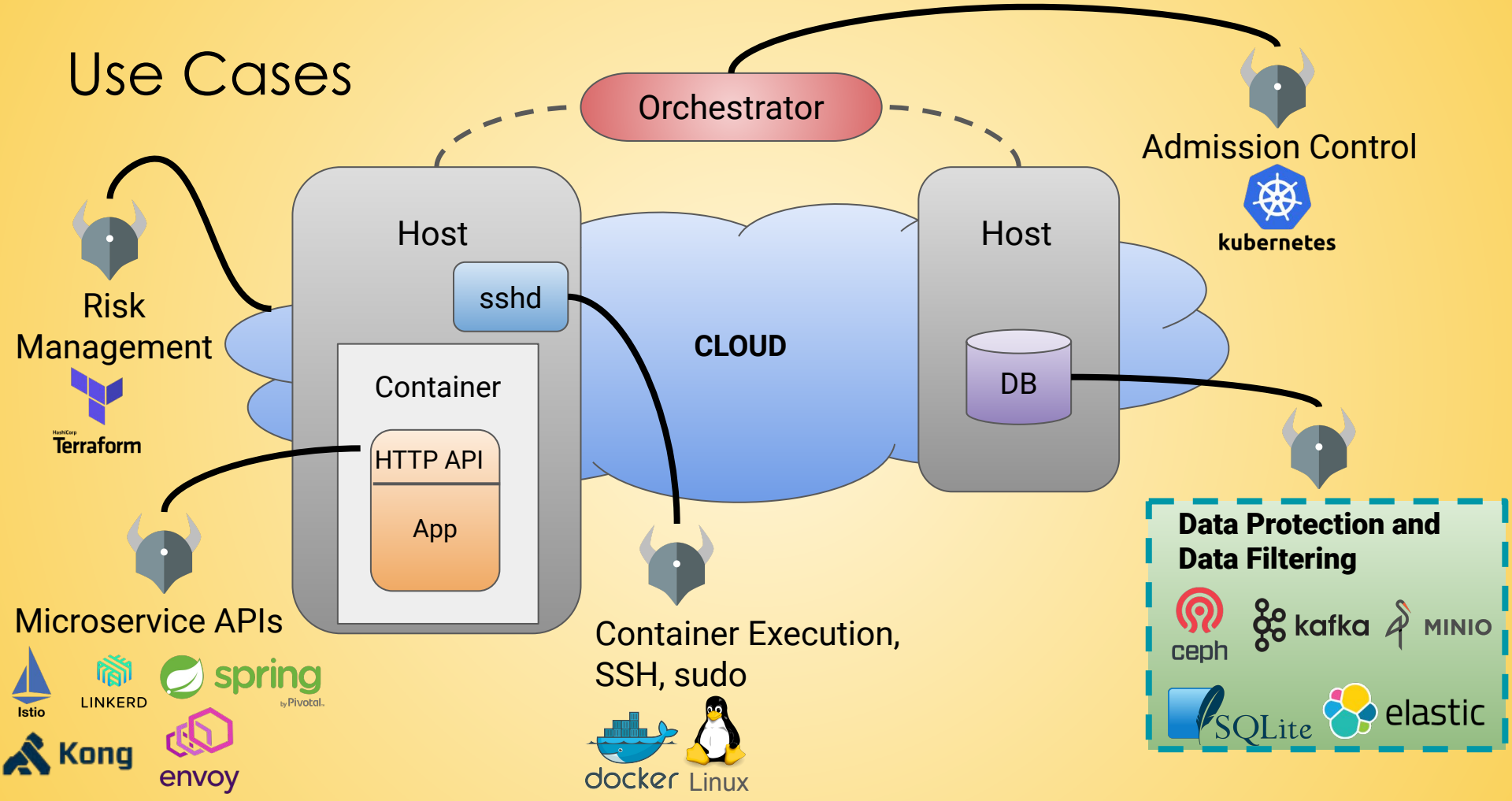
3 Steps to OPA 🎉

Step 1: Clone OPA Repo

Step 2: Build OPA binary

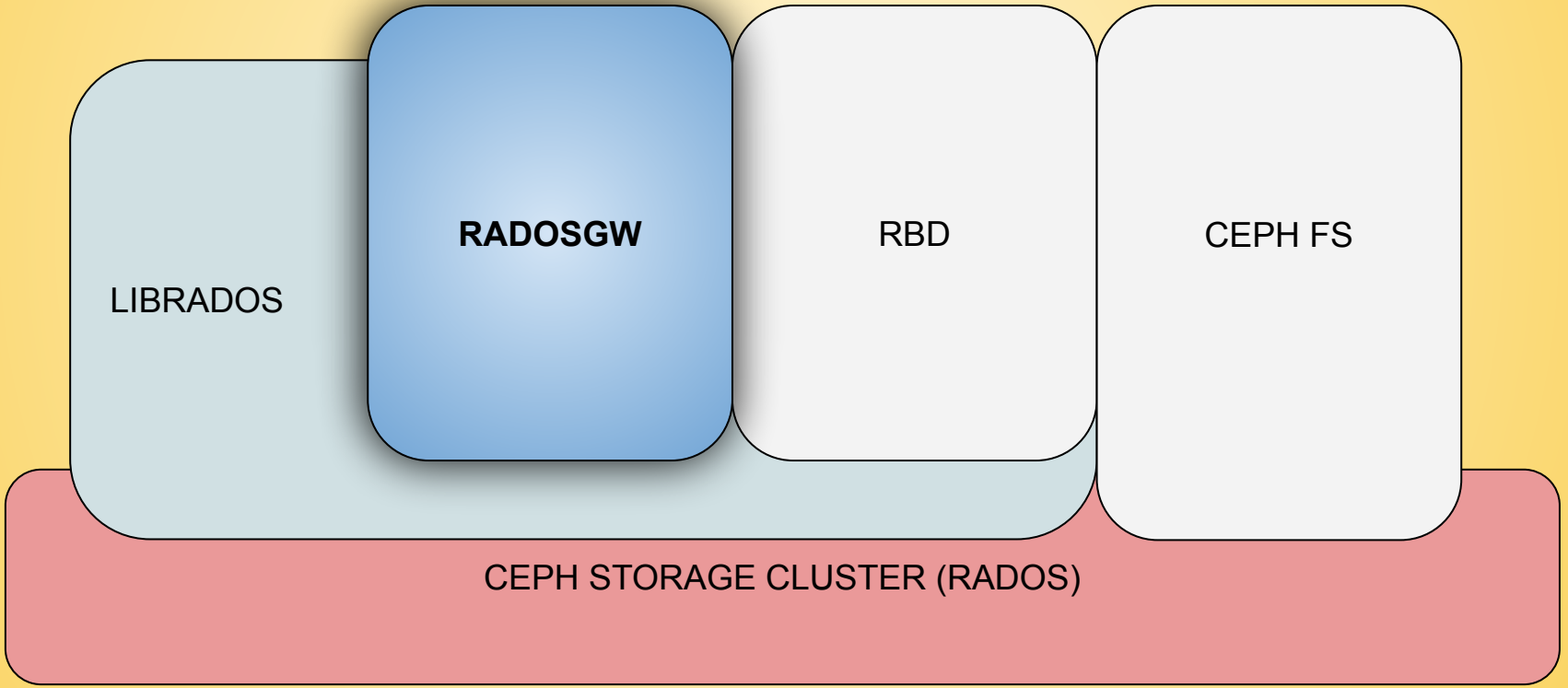
Step 3: Execute OPA binary

Use Cases

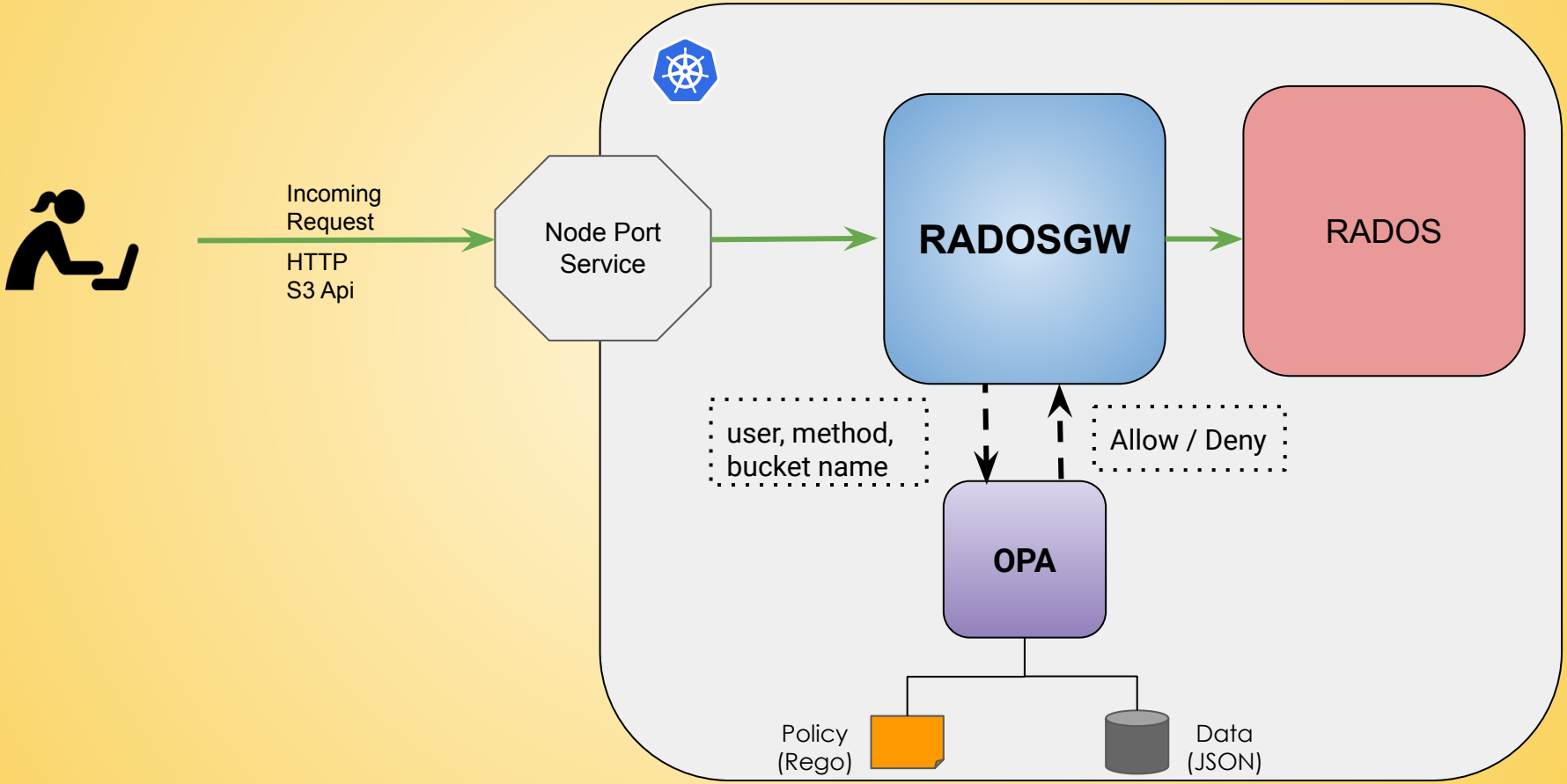

Orchestrator
Admission Control
kubernetes
Risk Management
Terraform
Host
sshd
Container
HTTP API
App
CLOUD
Host
DB
Microservice APIs
Istio
LINKERD
spring
by Pivotal
Kong
envoy
Container Execution, SSH, sudo
docker
Linux
Data Protection and Data Filtering
ceph
kafka
MINIO
SQLite
elastic

OPA Use Case: Ceph Data Protection

Ceph Architecture



Ceph Data Protection: Setup



Example policy

"Users can access only those buckets belonging to the same geographical region as them."

Demo: Ceph Data Protection

https://katacoda.com/styra

Data is King !

- Pervasive
- Abundant
- Customer Experience
- Revenue Growth

- Cyber Attacks
- Breaches
- Fines
- Loss of Customer Trust

Thank You!

styra Booth S20

openpolicyagent.org

slack.openpolicyagent.org

github.com/open-policy-agent/opa