

# Observability as Code for Al **Apps with New Relic and** Pulumi

**Piers Karsenbarg** Pulumi

**Harry Kimpel** Senior Solutions Engineer Developer Relations Engineer **New Relic** 

## Housekeeping

- Be chatty in the Chat tab
- Ask questions in the Q&A tab
- Slides+Code are in the Handouts tab
- YouTube recording link will be sent in a follow-up email

## **Agenda**

- About Pulumi
- Pulumi ESC
- About New Relic
- New Relic Al monitoring
- Demo
- Questions and Answers



# **About Pulumi**





### What is Pulumi?



modern infrastructure as code (IaC)

A文

uses familiar languages and tools



works with any major cloud

Open Source

Apache 2.0

Free Tier Managed SaaS



### Pulumi Infrastructure as Code



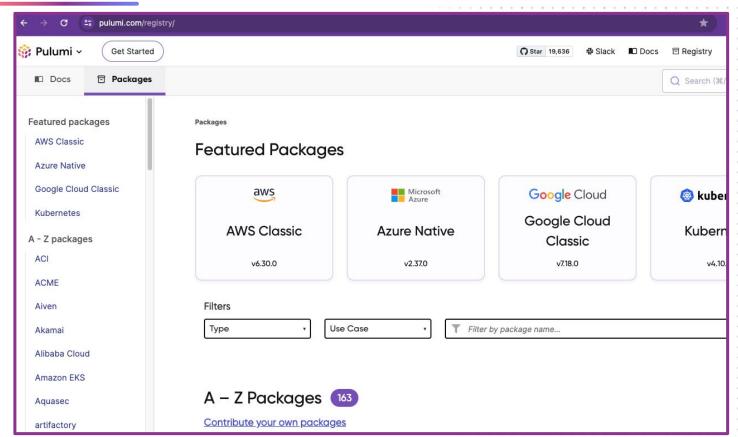




**160+** Integrations



## **Pulumi registry**





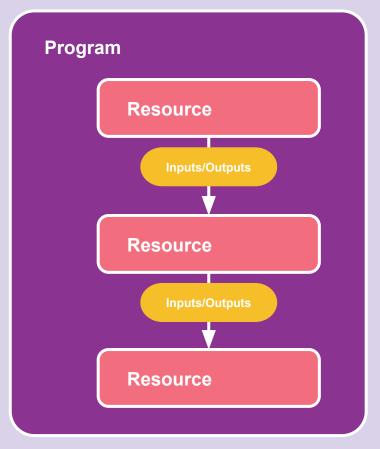
### Pulumi vs. others

**Pulumi** General purpose Language (multiple) **Resources** Any cloud **Deployment V**Local CLI Model ✓ Built-in **Secrets** Handling (state and config)

Terraform	CDK	CloudFormation
HCL	General purpose (multiple)	YAML
✓Any cloud	AWS only	AWS only
✓ Local CLI	CloudFormation Service	CloudFormation Service
Requires Vault	Not built-in (can use AWS KMS)	Not built-in (can use AWS KMS)



#### **Project**



Stack (DEV)

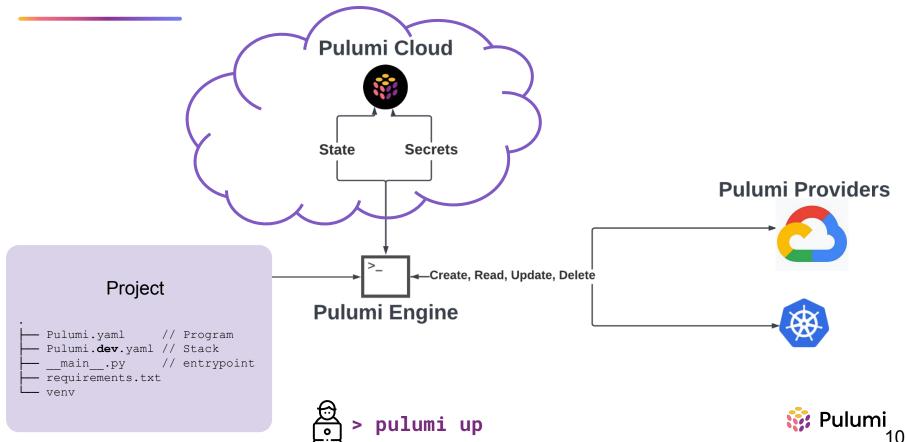
Stack (QA)

Stack (PROD)

## Pulumi Program Model



## Behind the scenes



### **Pulumi ESC**

#### TAME YOUR SECRETS AND CONFIGURATION SPRAWL

## **Encrypt**

all sensitive data in-transit and at-rest.

### **Centralize**

management of cloud configuration and secrets at scale.

## **Equip**

teams to securely store API keys, tokens, credentials, certificates, etc

## **Enforce**

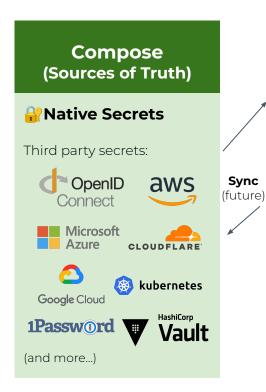
security best practices by-default, such as short-lived credentials.

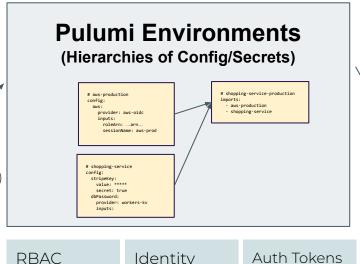


### **Pulumi ESC**

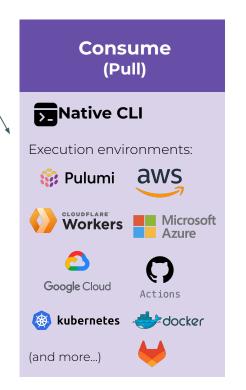


#### **EXTENSIBLE ARCHITECTURE - SOURCES AND SINKS**





RBAC	Identity	Auth Tokens
Encryption	Audit Logs	Managed or OSS
Policy	Versioning	Web UX





New Relic All-in-one Observability Platform





The world's most powerful telemetry platform – at a low, per-GB cost.



High capacity platform with 4k+ queries/second (QPS); storage, compute, and services to satisfy most use cases.



#### **Lightning-fast queries**

Highly responsive platform to give users responses fast – 95% of requests achieved in less than 880 ms.



#### Massive scale

Massive surface area of our telemetry data platform, with 200 billion data points scanned every second.



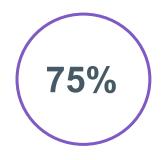
### **Market Trends**



\*Organizations currently in explorative mode with generative AI



\*\*Enterprise applications will have conversational AI embedded by 2024



\*\*\*Consumers are concerned about the misinformation from Al



Source

\* Gartner Poll

\*\* Gartner Insights Generative Al

\*\*\* Forbes Al Statistics



## Challenges monitoring AI applications



#### Increased complexity

Al tech stacks include new components like LLMs, vector data stores and orchestration frameworks.



#### **New telemetry**

Al applications introduce new challenges with quality of responses(Bias, Hallucination and Toxicity) and cost of managing models



#### **Security & Compliance**

Al regulations are still emerging and customers don't trust sharing proprietary information with AI/LLM providers

Customers need new operational capabilities to be able to monitor/debug/fix AI/LLM latency, cost and response quality.



## **New Relic Al Monitoring**

Industry's first end-to-end APM solution that empowers engineers to build and run AI applications confidently with in-depth insights across the AI application stack to optimize performance, quality and cost



### Compliance: security & data protection measures

#### Disable AIM

- disable AIM during the agent installation
- see <u>Customize</u> agent for Al monitoring

#### Disable content

- enable AIM but disable sending message contents, where sensitive data might exists
- see <u>Customize</u> agent for Al monitoring

#### **Drop filters**

- enable AIM and messages content
- create drop filters to remove any data which match potential sensitive data (e.g. phone, email etc.)
- based on regex
- see <u>Drop sensitive</u> <u>data with drop</u> <u>filters</u>



## **Learn more!**

Stay tuned on upcoming workshops and in-person events!

pulumi.com/resources



# Thank you









