



Google Next 23'

nino.chang@swag.live

Google Next '23'

| Outline

- *Duet AI* integration to GCP products
- *VertexAI* updates
- *Infrustructures* updates
- Sessions sharing
 - 1. Build an *AIOps* platform at enterprise scale with Google Cloud
 - 2. Best practices for *DevOps velocity* and *security* on Google Cloud
 - 3. Running large-scale *ML* on *GKE*
 - 4. *Prompt engineering*: Getting the skill your team needs next

puretea

| Features

- Assisted development
- Assisted operations
- Assisted data
- Assisted no code development
- Assisted security

| Assisted Development

- (2023) vscode¹ / cloudshell
- (Future) PyCharm / Intel / IntelliJ IDEA / GoLand / WebStorm
- Real time code completion/generation and autofix suggestions.
- Providing code license attribution
- Chat²
 - Explain code
 - Suggest a test plan

¹[vscode integration document](#)

²[chat with Duet AI document](#)

| Demo VScode

| Assisted Operations

- Get help where you are
- Transition from your question to cloud shell command
- Contextualize log explanations
- Looker studio insight generation
- meet
 - realtime translate/transcript
 - summarize meet content textually

SQL Instances

+ CREATE INSTANCE MIGRATE DATA ⋮

⚠️ 1 instance at or nearing capacity. Prevent SQL instance from running out of disk space and going offline by manually or automatically creating storage space.

VIEW ALL

ℹ️ You have 328 recommendations across 315 instances for this project.

VIEW ALL

Filter Enter property name or value ? ☰

Instance ID ? ↑	Recommendations	Maintenance	Type	Actions
<input type="checkbox"/> a-g-test		-	P	⋮
<input type="checkbox"/> a-garden-prod-replica-clone-2		-	M	⋮
<input type="checkbox"/> a-garden-prod-replica-clone-2-replica		-	M	⋮
<input type="checkbox"/> a11y-test		-	M	⋮
<input type="checkbox"/> a1233	Configure SSL	-	M	⋮
<input type="checkbox"/> a1233-replica-2	Configure SSL	-	M	⋮
<input type="checkbox"/> a1233-clone	Configure SSL	-	M	⋮
<input type="checkbox"/> a1233-swag	Configure SSL	-	M	⋮

How can I reduce Cloud SQL costs?

Here are some strategies for optimizing the costs of Cloud SQL instances:

- Use the right Cloud SQL instance type based on your performance requirements, such a general-purpose instance or a memory-optimized instance.
- Identify and turn off unused instances. If you have instances that are not being used, consider turning them off to avoid unnecessary charges. Example commands to retrieve and stop idle instances:

```
$ gcloud recommender recommendations list \ --proj
```

Use code with caution. [Learn more](#)

- Use long term [commitment discounts](#). Cloud SQL committed use discounts give you a 25% discount off of on-demand pricing for a 1-year commitment and a 52% for a 3 year commitment.

Rate this answer: 👍 👎

⌄ Show related content

Enter a prompt here ❯

Google Cloud stofe-storage Search (/) for resour... Search Duet AI PREVIEW

Logs Explorer REFINE SCOPE Project SHARE LINK LEARN

Query Recent (2) Saved (0) Suggested (22) Library Clear query Save Stream logs Run query

Last 1 hour Search all fields Resource Log name Severity Show query

Log fields Histogram Actions

Log fields <> Histogram

SEARCH fields and values

RESOURCE TYPE

Resource Type	Count
VM Instance	117,734
Kubernetes Cluster	69,687
Kubernetes Node	51,680
Cloud Dataproc Cluster	28,401
Cloud SQL Database	18,963
Dataflow Step	8,370
Dataproc Metastore Service	5,161
Node Group	2,596
AlloyDB instance	1,292
Cloud Pub/Sub subscription	877

Show more

Histogram 20K Jun 1, 11:08:00 AM 11:30 AM Jun 1, 12:08:30 PM

Query 306,228 log entries Find in results Correlate by Download

SEVERITY TIMESTAMP PDT SUMMARY EDIT

2023-06-01 12:07:50.879 PDT gke-asia-northeast3-dele-default-pool-3b93a0go/informers/factory.go:135: Failed to list *v1.Endpoints: dial https://34.64.239.122/api/v1/endpoints? refused

gke-asia-northeast3-dele-default-pool-3b93a0go/informers/factory.go:135: Failed to list *v1.Endpoints: dial https://34.64.239.122/api/v1/endpoints? refused

Hide log summary Expand nested fields Copy to clipboard Copy link

Similar entries

insertId: "a7wwbv1tc7h0ytfz"

jsonPayload: { message: "k8s.io/client-go/informers/factory.go:135: Failed to list *v1.Endpoints: Get https://34.64.239.122/api/v1/endpoints?" }

Explain this log

Additional prompt details

The log you provided is an error log from the kube-proxy service.

This error means that the kube-proxy service was unable to connect to the Kubernetes API server at the address `https://34.64.239.122/api/v1/endpoints`. This could be due to a number of reasons, such as the Kubernetes API server being down, the kube-proxy service not being able to resolve the address, or the kube-proxy service not being able to connect to the Kubernetes API server over HTTPS.

To resolve this error, you will need to first check that the Kubernetes API server is up and running. You can do this by running the following command:

kubectl get nodes

Use code with caution. [Learn more](#)

If the Kubernetes API server is up and running, you will see a list of nodes returned. If the Kubernetes API server is not up and running, you will see an error message.

Enter a prompt here

| Assisted Data

- SQL Completion
- SQL Generation
- SQL Generation-Iteration
- SQL Explainability

| Demo bigquery suggestion.

- -- Find Nino's salary

- -- Find Nino's salary from **playground.employees**

- -- Find all employee from **playground.employees** who's salary higher than 10

- -- Use query above filter out member come from department_id 4

| Assisted no code development

- with App sheet³
- Build app conversationally

³[app sheet](#)

| Assisted security

Duet AI integrated with

- Madiant
- Chronicle

VerteXAN

| Vision - Generation

- Digital watermarking
- image styling traing and apply

Vision

EXPORT RESET HISTORY PROMPT GUIDE

Parameters >

We want your [feedback](#)

Settings

Number of results: 4

Negative prompt: Define what you don't want to see

Style

🚛 🌴 🌴

GENERATE EDIT CAPTION VISUAL Q & A VERIFY

Freight truck high quality

Add details like the context and style in your prompt. Model may output inaccurate or offensive imagery that doesn't represent Google's views. [Report inappropriate content](#)

The screenshot shows the Google Vision interface with a sidebar of icons. The main area displays four generated images: a blue semi-truck, a white semi-truck, a red semi-truck, and a close-up of a truck's front end. Below these are buttons for GENERATE, EDIT, CAPTION, VISUAL Q & A, and VERIFY. A text input field contains "Freight truck high quality". On the right, there are five additional images: a field of tall grass, a white sports car, a sunset over mountains, a boat in the water, and a beach scene.

Vision

EXPORT RESET HISTORY PROMPT GUIDE

The screenshot shows the Google Vision interface with a sidebar of icons. The main area displays three generated images of a semi-truck driving on a road at sunset. Below these are buttons for EXPORT, RESET, HISTORY, and PROMPT GUIDE. A small thumbnail of a truck image is visible at the bottom.

| LLMs

- variations of model ready to use as fundamental.
 - Llama2 (2023/7)
 - Code Llama (2023/8)
 - Palm (2023/3)
 - up to 47 basic models.

| Infrastructures updates

| Infra

- A3 VMs + Nvidia H100 GPU⁴
 - Deploy A3 on VertexAI or GKE
 - 10x more network bandwidth compared to our A2 VM (GPU-to-GPU data transfer)
- Cloud TPU v5e
- GKE enterprise⁵

⁴[A3](#)

⁵[GKE version Comparison](#)

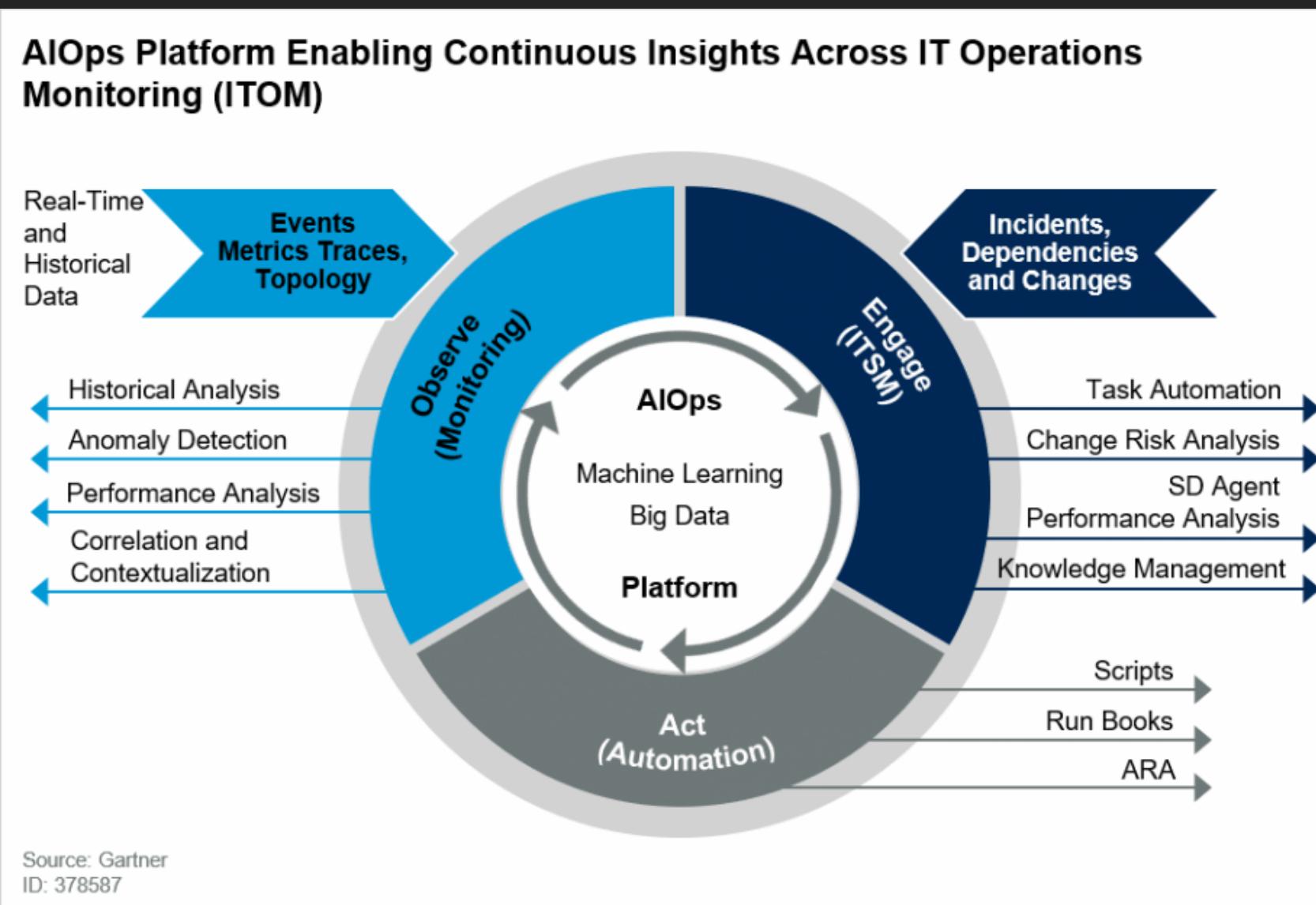
| Sessions Sharing

| Build an **AIOps** platform at enterprise scale with Google Cloud

| Solve Simple problem -

Need to triage an application outage

- Did something change recently ?
- What systems are involved ?
- How to reduce finger-pointing and need for war rooms to solve the issue



The application of AI capabilities, such as natural language processing and machine learning models, to *automate* and streamline operational workflows.

| Data from inside

- Machine: The logging of performance statistics, storage, CPU, memory.
- Application: Applications logs
- Network / Traffic: Is traffic spiking from established norms?
- Security: Access logs, Change logs, firewall events.
- Synthetic: Are synthetic transactions responding as expected ?

| Data from outside

- Human Sentiment: Customers express particular service is down as a tweet on twitter.
- Incident Tickets: How do miscategorized tickets impact *Actual Business Impact*
- Weather / Seismic Alerts: Area infrastructure been impacted.

Infrastructure Intelligence Platform - Vision

Collect

Transform

Analyze

Visualize

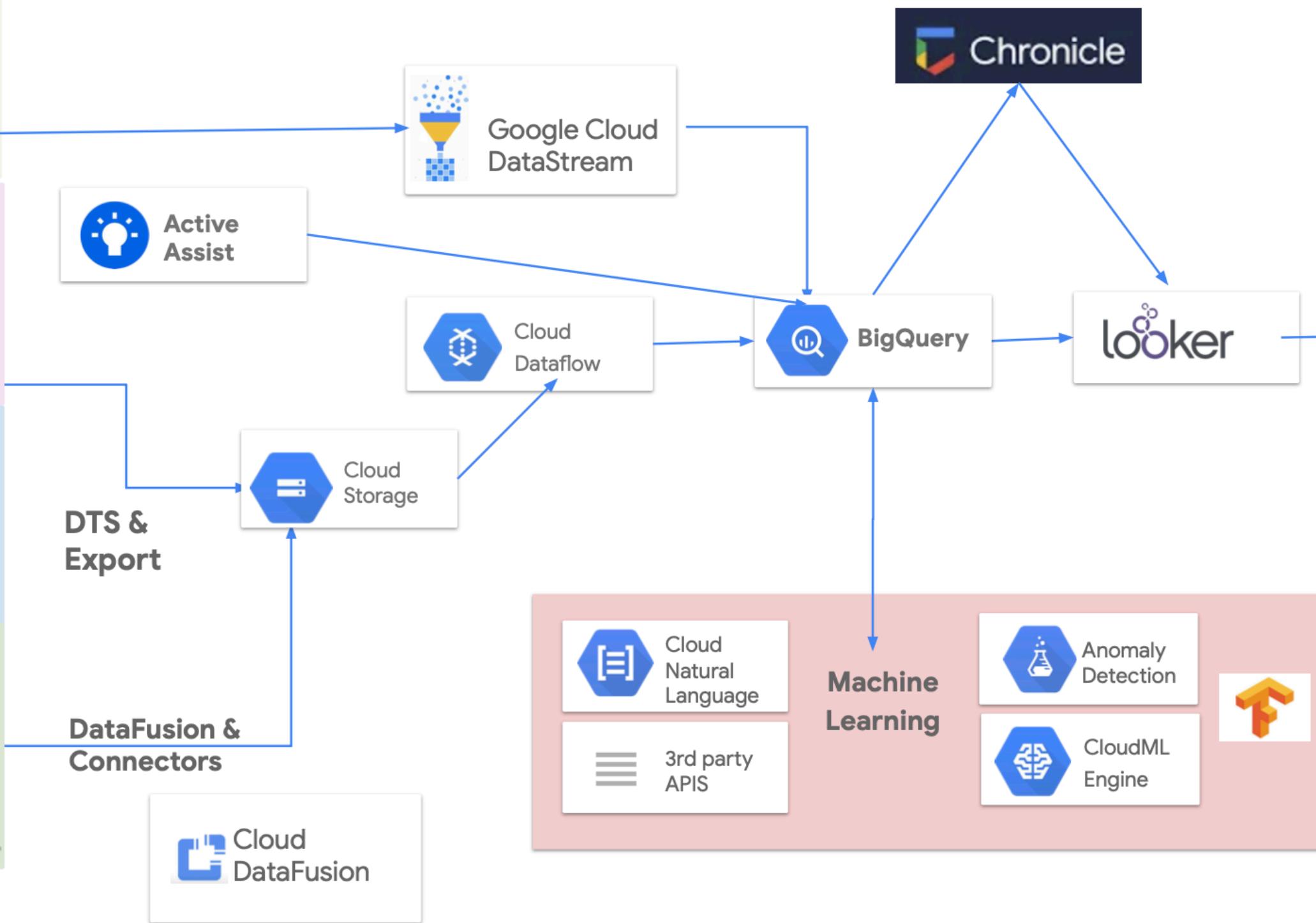
Take Action

- Machine
- Network
- Security
- Synthetic
- Active Assist

- Application
- Agent
- Security
- Synthetic

- Weather
- Seismic
- Traffic
- Other

- Social
- Sentiment
- Support Tickets
- Internal chatter



Active Assist Insights & Recommendations



Cost

Manage your costs wisely



Reliability

Deliver highly available and reliable services



Security

Mitigate your security risks proactively



Manageability

Spend less time managing your cloud



Performance

Maximize the performance of your systems



Sustainability

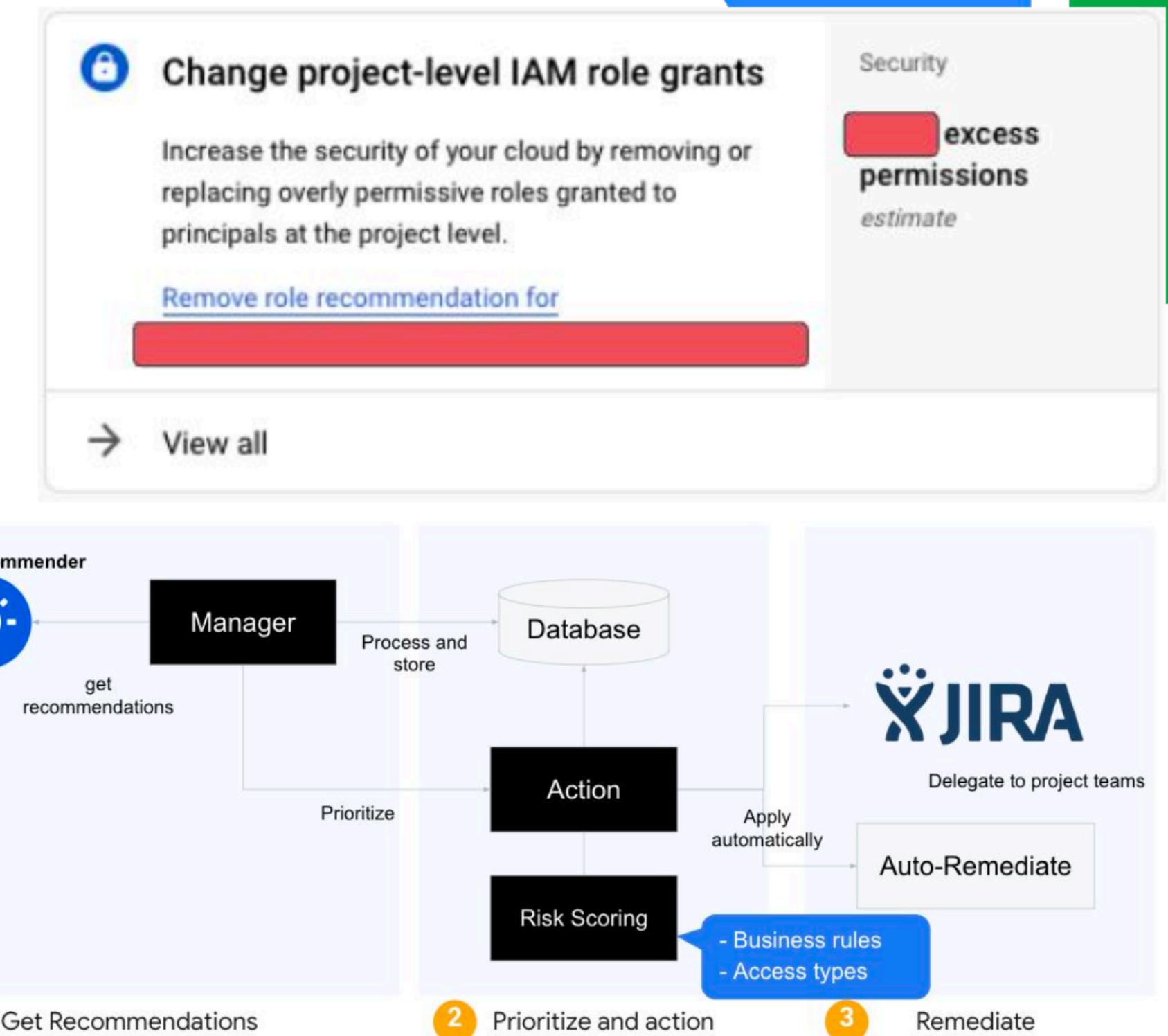
Reduce the carbon footprint of your workloads

| Uber's 3 use case of **Active Assist**⁶

⁶ [Active assist](#)

IAM Recommender

- IAM Roles may be over-granted, creating security risk
- Finding and reducing over-granted permissions is a complex task
- Solution: automation utilizing Active Assist IAM Recommender
- Assigns Jira tickets to project owners with recommendations and validates fix

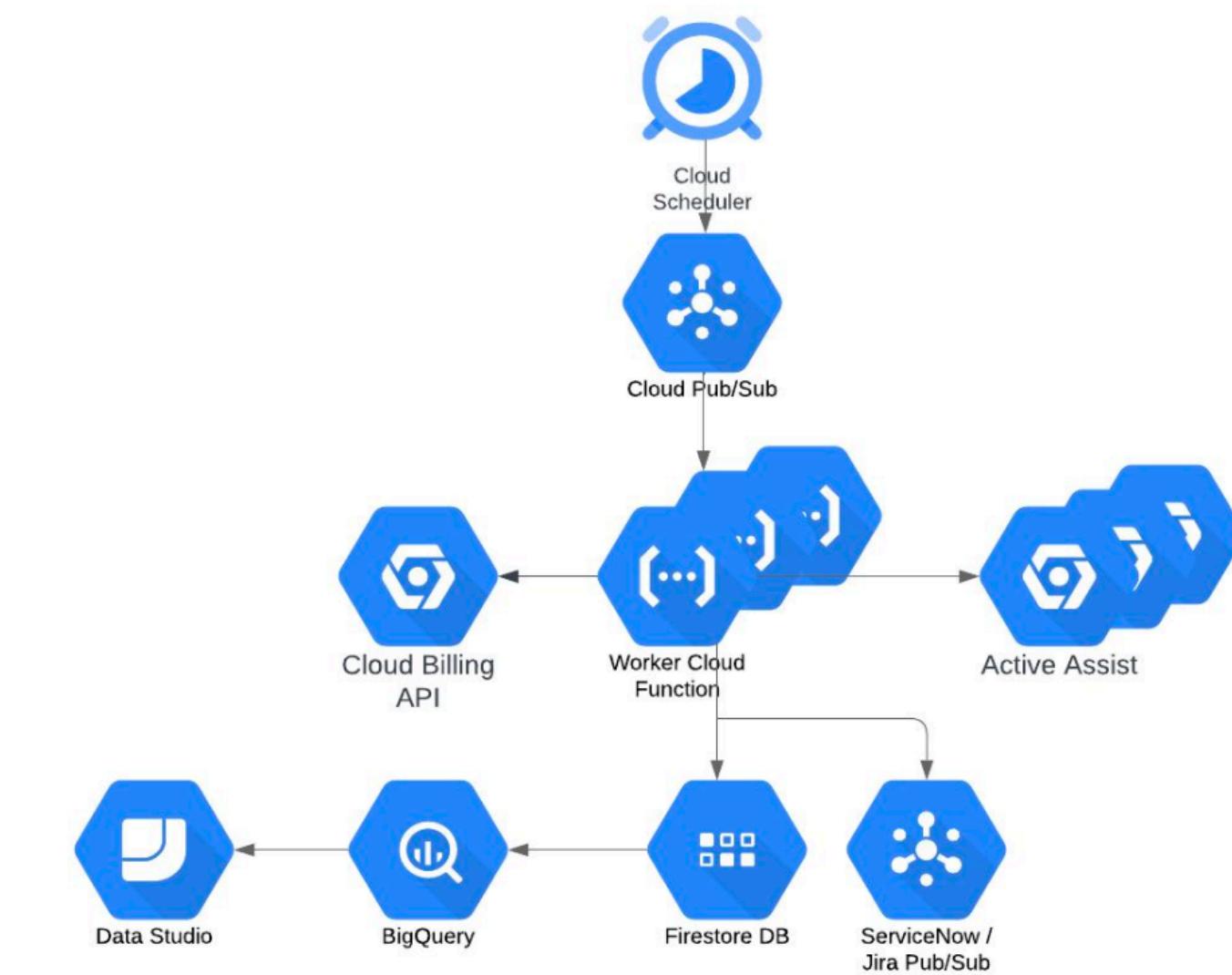


swag

Unattended Projects Recommender

GCP Project Lifecycle Automation

- Monitors inactive and orphaned GCP Projects utilizing Active Assist
- From inefficient logging queries to Google Active Assist
- Reduction in Attack Surface, Cost and CO₂



Compute Engine Recommender

GCP Compute Right-scale Automation

- Monitors inactive and underutilized Compute VM, Disks and Images
- Assigns tickets to resource owners with metrics-based recommendations
- Reduction in Attack Surface, Cost and CO2

The screenshot shows two main sections of the GCP Compute Engine Recommender interface.

Unused Compute Engine resources: This section displays a cost savings estimate of \$15.92/month for deleting a disk. It includes a link to "Delete disk to save \$13.55/month" and a "View all" button.

Cost savings: \$15.92/month estimate

Resize instance: This section shows a recommendation to switch from the current machine type (n1-standard-1) to a new machine type (g1-small). It includes a note about low memory utilization, a "Learn more" link, and a "Customize" button. A warning message states that changing the machine type requires stopping and starting the instance, which may cause data loss.

Resizing options:

- Current machine type: n1-standard-1 (1 vCPU, 3.75 GB memory)
- New machine type: g1-small (1 vCPU, 1.7 GB memory) Recommended

Recommendations:

- Consider installing the Monitoring Agent for more precise recommendations [Learn more](#)
- To change the machine type, Compute Engine needs to stop and start this VM instance. Stopping and starting the machine may cause resources, such as local SSDs and ephemeral IP addresses, to be lost.

Buttons at the bottom: CANCEL, DISMISS RECOMMENDATION, APPLY



| GitOps tools come in 2024.

| Best practices for DevOps velocity and security on Google Cloud

Measuring Software Delivery Performance - The four keys



Lead time for changes

How long does it take to go from code committed to code successfully running in production?



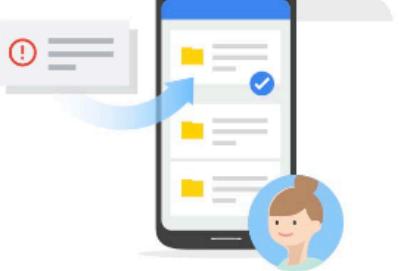
Deployment frequency

How often does your organization deploy code to production or release it to end users?



Change fail rate

What percentage of changes to production or releases to users result in degraded service?



Mean time to restore service

How long does it generally take to restore service when a service incident or a defect that impacts users occurs?

Throughput

Stability

swag

- Improve anything except for recover time (since roll back is easy in cloud)
- Use canary deployment gke stragety
- Work on stability first and then speed.
- Run unit test in cloud deploy, if test failed don't merge prs.
- Cloud deploy do port forward for you

| Demo cloud build increase

- velocity
- security

| Cloud build trigger build and test on pull request

steps:

```
# Build and tag using commit sha
- name: 'gcr.io/cloud-builders/docker'
  args: ['build', '.', '-t', 'asia-east1-docker.pkg.dev/$PROJECT_ID/pop-stats/pop-stats:${COMMIT_SHA}', '-f', 'Dockerfile']
  dir: 'app'

# Run api tests
- name: 'asia-east1-docker.pkg.dev/$PROJECT_ID/pop-stats/pop-stats:${COMMIT_SHA}'
  entrypoint: python
  args: ["api_tests.py"]
  dir: 'app'
```

| Cloud build trigger when main been merged

```
apiVersion: deploy.cloud.google.com/v1
kind: DeliveryPipeline
metadata:
  name: pop-stats-pipeline
description: pop-stats application delivery pipeline
serialPipeline:
  stages:
    - targetId: staging
      profiles:
        - staging
    - targetId: prod
      profiles:
        - prod
strategy:
  canary:
    runtimeConfig:
      kubernetes:
        serviceNetworking:
          service: "pop-stats"
          deployment: "pop-stats"
canaryDeployment:
  percentages: [5]
  verify: true
```

| Cloud build trigger when main been merged

```
apiVersion: deploy.cloud.google.com/v1
kind: DeliveryPipeline
metadata:
  name: pop-stats-pipeline
description: pop-stats application delivery pipeline
serialPipeline:
  stages:
    - targetId: staging
      profiles:
        - staging
    - targetId: prod
      profiles:
        - prod
strategy:
  canary:
    runtimeConfig:
      kubernetes:
        serviceNetworking:
          service: "pop-stats"
          deployment: "pop-stats"
canaryDeployment:
  percentages: [5]
  verify: true
```

Target staging

```
apiVersion: deploy.cloud.google.com/v1
kind: Target
metadata:
  name: staging
description: staging cluster
gke:
  cluster: projects/swag-nino-chang/locations/us-central1/clusters/stagingcluster
```

| Target prod

```
apiVersion: deploy.cloud.google.com/v1
kind: Target
metadata:
  name: prod
description: prod multi-target
requireApproval: true
multiTarget:
  targetIds: [prod1, prod2, prod3]
```

| Target prod

```
apiVersion: deploy.cloud.google.com/v1
kind: Target
metadata:
  name: prod
  description: prod multi-target
  requireApproval: true
multiTarget:
  targetIds: [prod1, prod2, prod3]
```

Gitlab x Google Cloud⁷



⁷ [Google partner with GitLab](#)

| Running large-scale machine learning (ML) on Google Kubernetes Engine (GKE)

0%
200



Tools and Libraries



Frameworks



Kueue: Kubernetes-native Job queuing



Autoscaling, Auto-provisioning, Scheduling, Lifecycle, Observability



Compute



TPU



GPU



Storage



Networking

swag

| Training

- TPU v4 and v5e (preview) available in GKE
- A3 available on GKE
- Use kueue manage jobs with in pods for different teams

| Serve large models affordably

- smaller container size using FasterTransformer
- splitting a model across many smaller GPUs
- Spot vs on-demand VMs

| Rapid pod startup saves money

- 3x faster application startup with *image streaming*
- Smaller containers using FasterTransformer and NVIDIA Triton
- Using GCS FUSE for pod startup latency (data intensive)

| Improve GPU utilization

- GPU time sharing up to 48 containers⁸
- Partition large GPU to many instance, reduce serving latency.
- Nvidia MPS, let application share GPU concurrently.

⁸[Time sharing GPU](#)

| Prompt engineering: Getting the skill your team needs next

| What's the most popular programing language in 2023 ?

English

| What kind of gen AI starts to see in 2023

- Use gen AI go through complex documents.
- Try to make data conversational, increase efficiency.
- Unstructured data, language, text, conversation.
- Audio/video other type of data.

| What's propt engineering and myth in it

- People think should talk like real communication but there's tricks you can use for talking to LLM.
- But with time, models gets less sensitive to which word you use.
- Try to provide detail information in question let model understand your question easier.
- Provide some example is a good idea for model to learn and response correct answer.

| How prompt engineering affect knowledge worker ?

- Emerge of new tool always affect all people
- Marketer used to use only news paper, now social media.
- Get it sonner, get benefit sonner.
- It kill jobs but also create new ones.
- Everyone need to learn make use of gen AI

| User should have good understanding on underlying model to be a better prompt engineer ?

- Need leading how to programming in english.
- Communication skill is more important.
- Context of application is more important than model, you need to evaluate the context in the end.
- How do we test "programming in english" if we treat it as software.
- Models give "believable" answer, not necessary "correct" answer.

| How can we better prepare our team for gen AI ?

- If you're not starting experimenting / play around with gen AI, you're behind
- Don't be afraid, keep experimenting.

| QA: Copyright ?

- Legal community still working on it.
- Still seek for balance knowledge sharing and keep secret inside.

| QA: What advice you will give younger age people that help career path to become prompt engineering ?

- Follow your passion, focus on what you like
- Many degree like Linguistic, are not aware it will come to work with engineering.
- Improve your communication skill

QA: What modalities are primed for the next big foundational model ?

- video model.
- multi modality model, input and output.

| Questions?