

Google Cloud

# Next '24

## Cloud Run: what's New



# Justin Mahood

Product Manager,  
Google Cloud



Google Cloud's serverless runtime

**Run applications  
fast and securely  
in a fully-managed environment.**



# Cloud Run



## Experience

### Simple

Demand as little as possible.

### Automated

Cloud Run takes care of a lot for you.

### Top satisfaction and usability scores

High satisfaction and task success.

### Developer productivity

Idiomatic to developers, deployment velocity.



## Runtime

### Capable

Run any container.

### On-demand

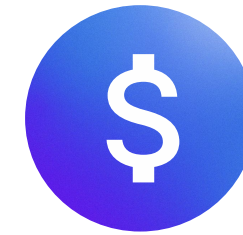
No pre-provisioning.

### Hyper-elastic

Fast automatic scaling, including to zero.

### No infrastructure management

No VM or cluster to manage.



## Pricing

Pay only when code is running, with a 100ms granularity.

- CPU
- Memory
- Requests (not always)

Perpetual monthly free tier

Committed Use Discounts

No flat fee!

*"if you don't use it, you don't pay for it"*



# Agenda

- 01 Simplifying App Development
- 02 Demos
- 03 Enterprise Ready
- 04 Customer Story: DZ BANK AG





# Simplifying App Development



# Volume Mounts

Preview

- Mount NFS or Cloud Storage Fuse filesystems as read-only or read-write
- Implemented through sidecar which handles kernel level file mount

Cloud Run | Create service

Container(s), Volumes, Networking, Security

CONTAINER(S) | **VOLUMES** | NETWORKING | SECURITY

Volumes

After creating a volume, navigate to the [Container\(s\) tab](#) to mount it to a container.

**New Volume**

Volume type  
Cloud Storage bucket

Volume name \*  
gcs-1

Bucket \*  
demo-cowsay-b120cf77 BROWSE

DONE

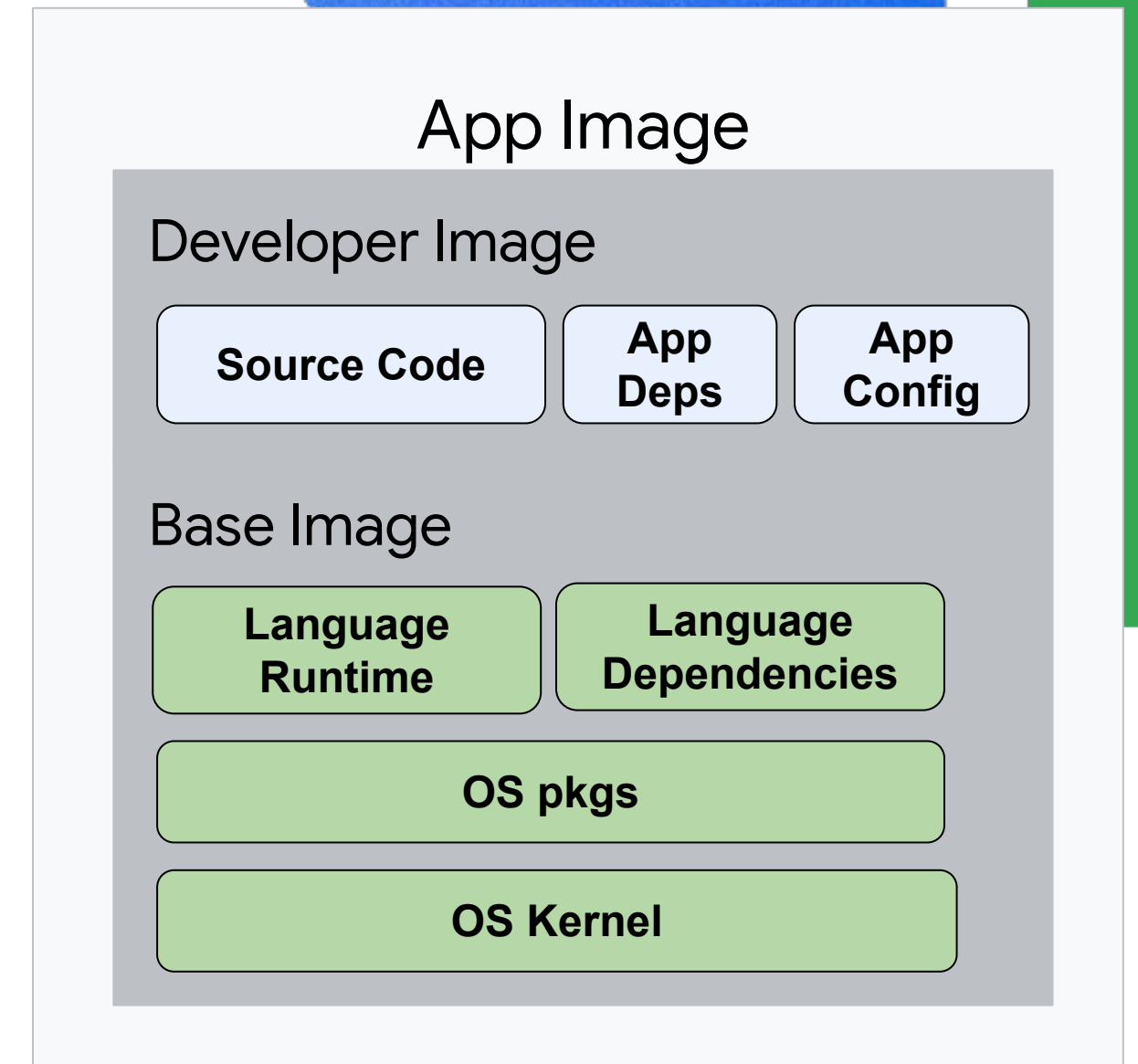
ADD VOLUME



# Automatic Security Updates

Private Preview

- Automatic base image updates on deployed images
- 0 downtime, 0 rebuild
- Target < 48hr patch time to security incidents



```
> gcloud run deploy myapp --source .  
--base-image=us-central1-docker.pkg.dev/serverless/google-22/go:latest
```



# Deterministic URL

Private Preview

- Access your Cloud Run service with a consistent URL scheme
- The leftmost DNS segment must be  $\leq 63$  characters long
- Coming soon: support for Private Google Access, and Private Service Connect
- Disable 'run.app' URL entirely if desired

https://**myservice**-**1234567**.**us-central1**.run.app

service name

project number

region

or

```
$ gcloud run deploy --no-default-url
```



# Gemini in Cloud Run Recommendations

Private Preview

- Surface recommendations proactively with banners and reactively with chat
- Aggregate recommendations for the most impact
- Summarize recommendations for easier impact analysis

**Insights generated by Gemini**

You can save approximately \$529.76 per month by changing the cpu allocation method for opt-cost2 and opt-cost1. opt-cost1 will have the most valuable impact of \$458.30 per month.

[Dismiss](#) [View Recommendations](#)

Name	Req/sec	Region	Authentication	Ingress	Recommendation	Last deployed	Deploy
opt-cost1	0	us-central1	Require authentication	All	COST SAVINGS	Feb 29, 2024	kritka
opt-cost2	0	us-central1	Require authentication	All	COST SAVINGS	Feb 29, 2024	kritka
opt-create-service-account	0	us-central1	Require authentication	All	SECURITY	Feb 29, 2024	kritka
opt-excess-permissions	0	us-central1	Require authentication	All		Feb 29, 2024	kritka
opt-performance	0.03	us-central1	Require authentication	All	2 RECOMMENDATIONS	Feb 29, 2024	kritka
opt-secrets	0	us-central1	Require authentication	All	3 RECOMMENDATIONS	Feb 29, 2024	kritka
service-dvpc	0	us-central1	Require authentication	All	PERFORMANCE	Feb 29, 2024	kritka
service-min-instances	0.03	us-central1	Require authentication	All	PERFORMANCE	Feb 29, 2024	kritka

**How can I lower the cost of my Cloud Run services?**

You can lower the cost of your Cloud Run services by changing CPU allocation for opt-cost1 and opt-cost2. The changes can save approximately \$529.76 per month.

Changing CPU Allocation to be always-on can reduce cost for certain types of workloads, for example, workloads with steady and slowly varying traffic. [Learn more](#).

- [Change CPU allocation](#) to save approximately \$458.30 per month for opt-cost1
- [Change CPU allocation](#) to save approximately \$71.46 per month for opt-cost2

Rate this answer: [👍](#) [👎](#)

Enter a prompt here

For best results use a detailed prompt. [Prompt guide](#)

Select a different project to use the GA version of Gemini.

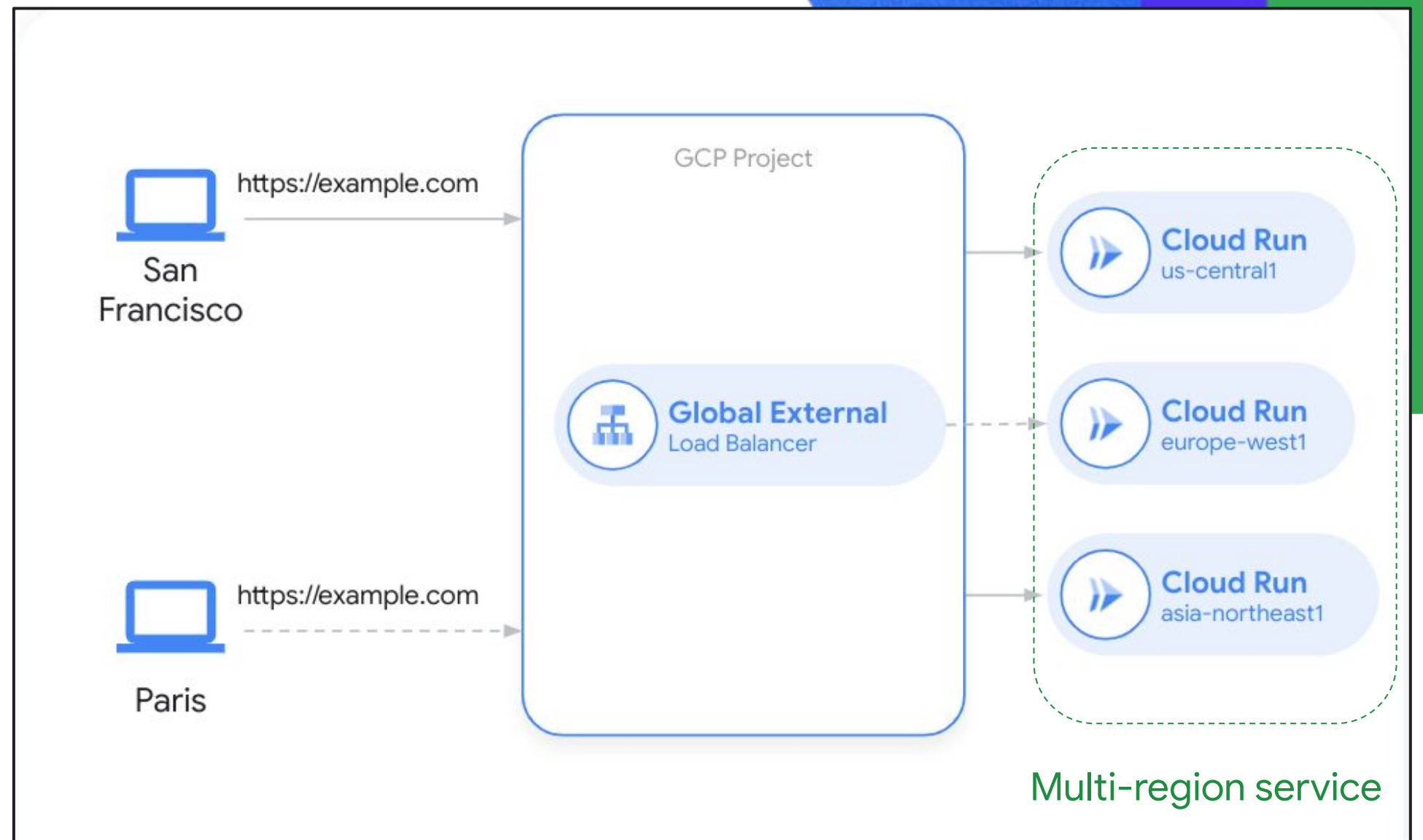
Google's use of Gemini for Cloud console is subject to Google's [policy for using generative AI tools internally](#), and the [Google Employee Privacy Policy](#)



# Multi-region services

Private Preview

- Deploy the same service to **multiple regions** with a **single command**
- Automatically expose a **global endpoint** that routes requests to the closest region



```
> gcloud run deploy myapp --region=us-central1,europe-west1,asia-east1  
--domain=example.com
```



# Demo



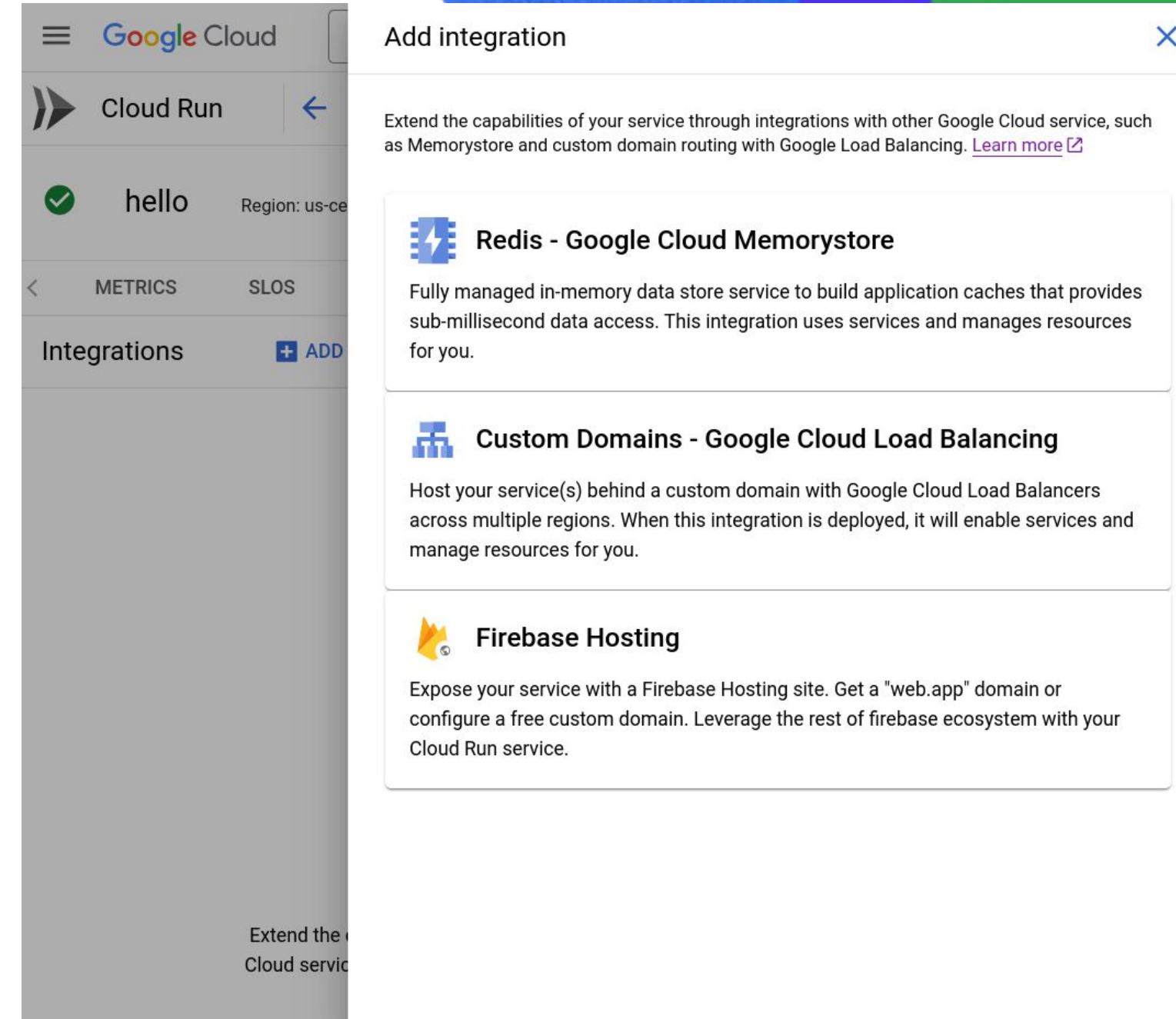
# Simplifying Application Development

**No application  
is just a container**



# Cloud Run Integrations

- One command for complex integrations
- Configures Cloud Run for you
- Easy access to all related resources





# New integrations

Preview

- Three new Cloud Run integrations: CloudSQL, Firestore, and VertexAI
- ‘1 click’ to automatically configure and connect GCP services to your Cloud Run workloads

## Add integration



Extend the capabilities of your service through integrations with other Google Cloud service, such as Memorystore and custom domain routing with Google Load Balancing. [Learn more](#)



### CloudSQL - Google Cloud SQL

Add MySQL, PostgreSQL, and SQL Server database services to your apps.



### Firestore

Use our flexible, scalable NoSQL cloud database, built on Google Cloud infrastructure, to store and sync data for client- and server-side development.

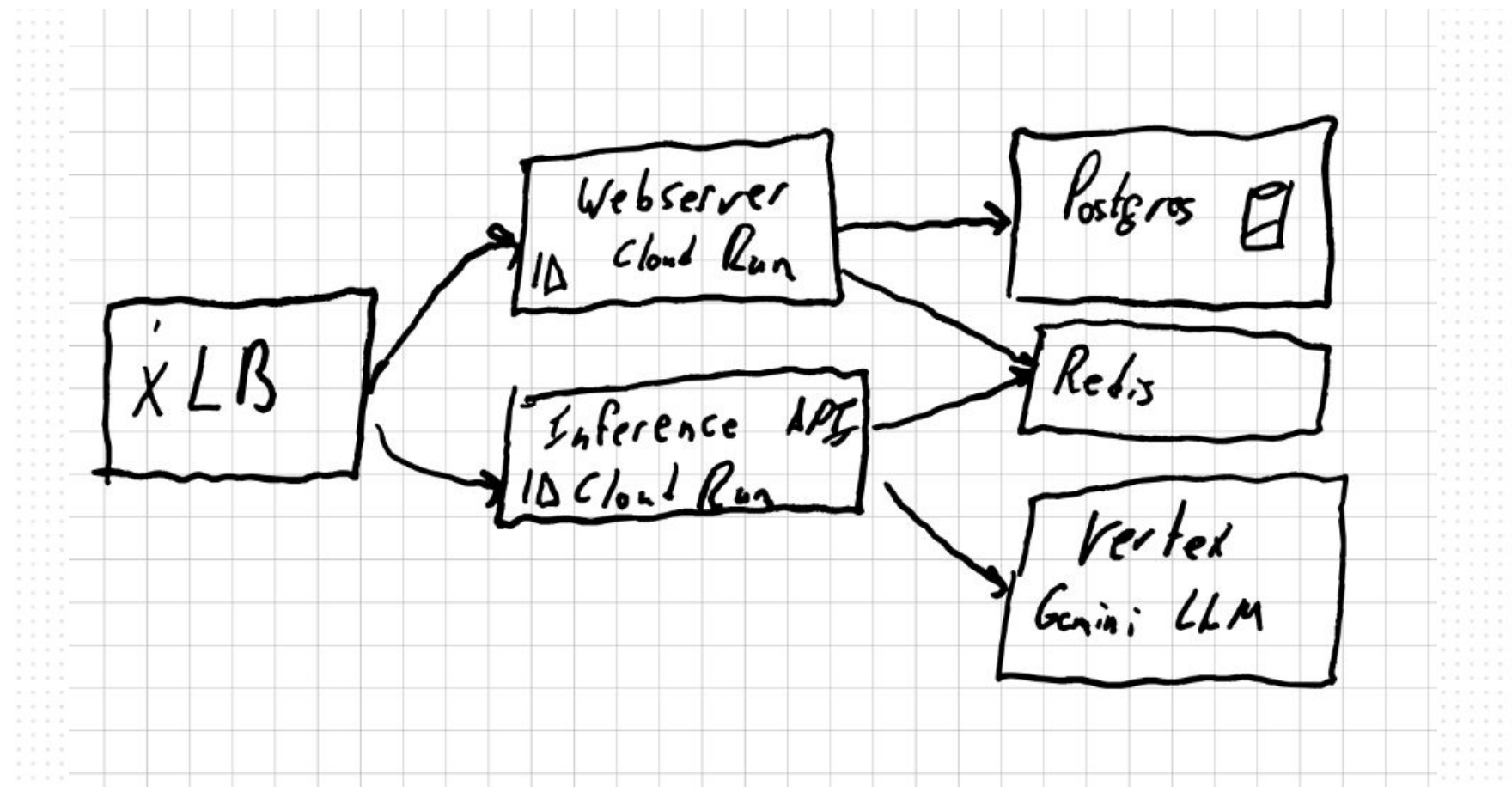


### Vertex AI - Generative AI

Configure access to Gemini, PaLM, Codey and more of Google's large generative models from your Cloud Run workloads



# Integrations were just the first step....





# Application canvas

Private Preview

- Simplify the creation of complex applications to a few clicks, and visualize in a diagram
- Use Gemini to edit or generate a **deployable** architecture using natural language

The screenshot displays the Google Cloud Application Canvas interface. At the top, it shows the navigation bar with 'Cloud Run', 'Manage Application' (with a 'PREVIEW' badge), and 'Region: us-east4'. Below the navigation bar, there are buttons for '+ ADD', 'UNDO', and 'REDO'. The main area is divided into two panels. The left panel contains a text input field with the prompt 'a cloud run service using an LLM and a vector database'. Below the input field, there are three buttons: 'Web app with database', 'MySQL instead of Postgres', and 'Add a cache'. A list of changes is shown: 'The following changes have been made to your architecture: • Added service/service1 using the gcr.io/cloudrun/hello:latest image. • Added vertex-genai/vertex-genai1. • Added cloudsql/cloudsql1 using PostgreSQL 14.' The right panel shows the configuration for the 'Service - Cloud Run' integration. It includes a 'Configure integration' header with a 'DELETE FROM APP' button. The service name is 'service1' and its status is 'Active'. Below this, there is an 'Integrations' section with a '+ ADD EXISTING RESOURCE' button. A table lists the integrated resources:

Resource type	Name	Status	Integration type	Action
Vertex AI - Generative AI	vertex-genai1	Not deployed	Backing Service	REMOVE
CloudSQL	cloudsql1	Not deployed	Backing Service	REMOVE

Below the integrations table, there is a 'Resources' section with a '+ ADD EXISTING RESOURCE' button. A table lists the resources:

Resource type	Status	Description
Cloud Run Service	Deployed	-

The bottom panel shows a diagram of the application architecture. It features a central 'Service service1' box with a 'Status: Draft' label. Two arrows point from this service to two other boxes: 'Vertex AI - Generative AI vertex-genai1' and 'CloudSQL cloudsql1', both with 'Status: Draft' labels.



# Demo



# Enterprise Ready



# Sridhar Venkatakrishnan

Engineering Manager,  
Google Cloud





# Enterprise workloads

Large enterprises  
have unique needs

- Advanced security and compliance needs
- Migrating from on-premises to cloud native containerized workloads
  - Slow starting containers (e.g. Java Spring Boot)
- Large scale
  - CPU and memory hungry
  - Many services and instances
  - Cost and performance sensitive
- Support complex network architectures spanning multiple domains

# Secure

## Access Control

IAM invoker permission

VPC Service Controls

Identity Aware Proxy

## Supply Chain

Secure Software Supply  
Chain insights in Console

Automatic base image  
updates (Private Preview)

Binary Authorization

## Encryption

Customer Managed  
Encryption Keys

Secret Manager integration

## Compliance

ISO, SOC, PCI

FedRAMP Moderate, NIST, IL2

HIPAA

Assured Workload

## Network Security

Ingress & egress controls

Firewall rules with Direct  
VPC Egress

Disable default run.app  
URL (Private Preview)

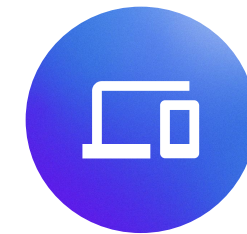


# Cost effective at scale



## Max Instances

Specify a maximum number of instances to limit the upper-bound of scaling



## Always on CPU

"Always-on CPU" is priced 25% lower than "throttled CPU"  
More economical for large scale workloads



## Committed Usage Discounts

Self-service Committed Use Discounts offering a 17% reduction in usage cost.



## Active Assist

Proactive recommendations to optimize pricing selection and commitment.



# Announcing: Direct VPC Egress GA

**Scalable** 2x throughput compared to VPC Connectors  
Lower latency  
Lower cost

**Compatible** All Cloud Run regions supported\*  
Cloud NAT\*  
Firewall rules logging and VPC flow logs\*

**Simple** Simple configuration  
Manage with org policy  
Instance limits are now a quota\*



# Announcing:

## Cloud Service Mesh with Cloud Run

Private Preview

### Benefits

- Traffic management
- Observability
- Security

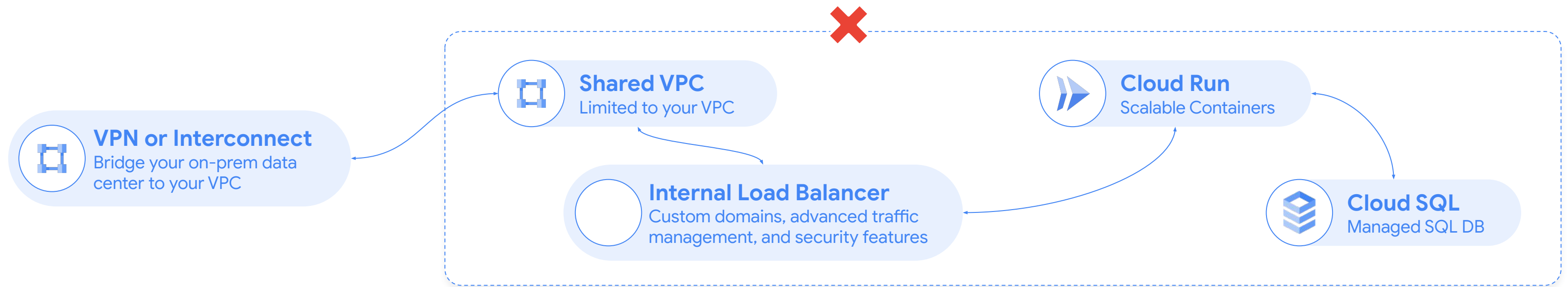
### Notable Cloud Run features

- Fully-managed data plane (Envoy sidecars)
- Friendly service names
- JWT token injection



# Integrated

Integrate easily with your existing networks and systems



(Optional) on-prem VM calling through **VPN or Interconnect**

Your **private shared VPC** may contain internal resources and users with a security boundary enforced at the network level

**Internal Load Balancer** gives you custom domains, advanced traffic management, and security features

**Cloud Run** will only accept requests from within your project or shared VPC network, and will prevent egress to any destination outside the VPC

Other Google Cloud resources within the VPC boundary are accessible



# Customer Story: DZ BANK AG



# Tim Harpe

Senior Cloud Engineer,  
DZ Bank

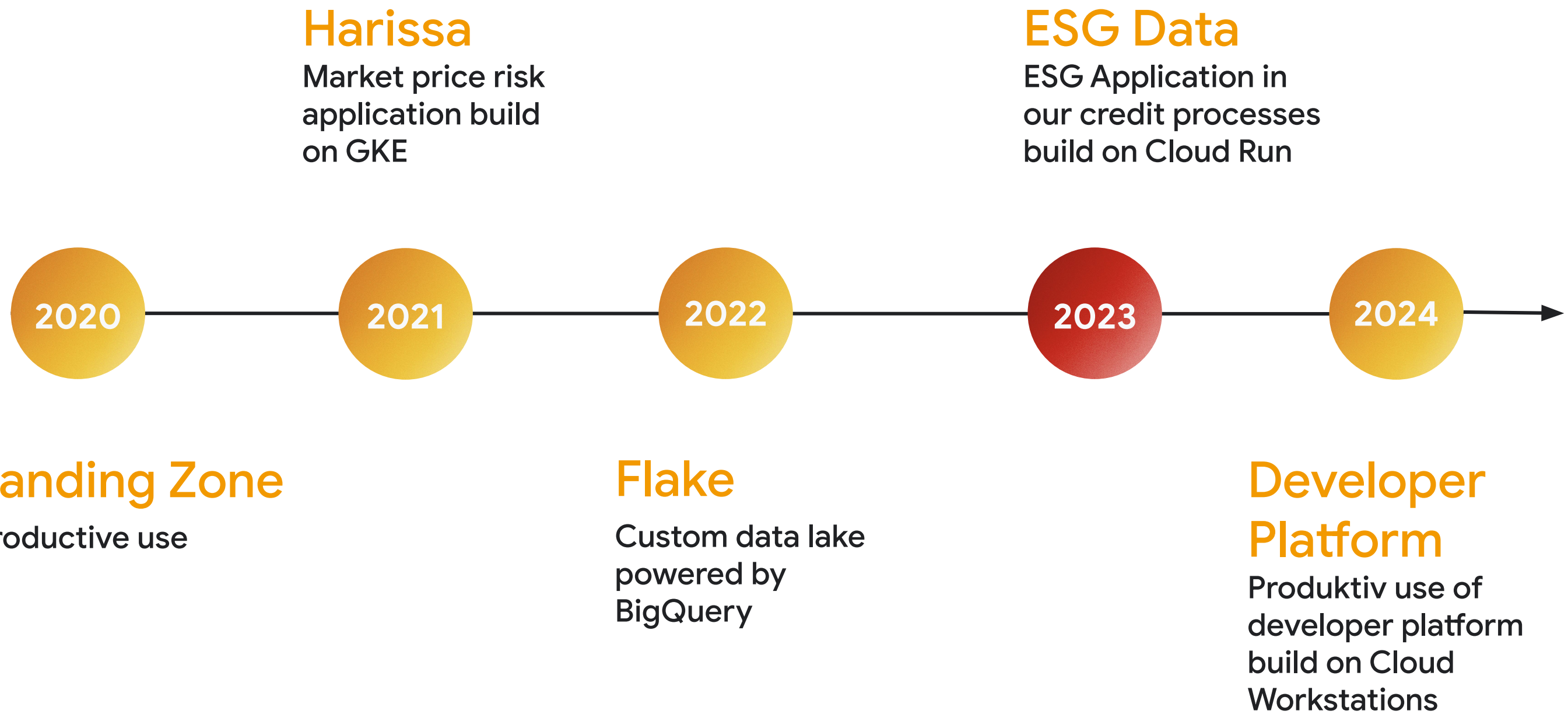






**Industry:** Financial Services  
**Country:** Germany 

 **Customer Award**

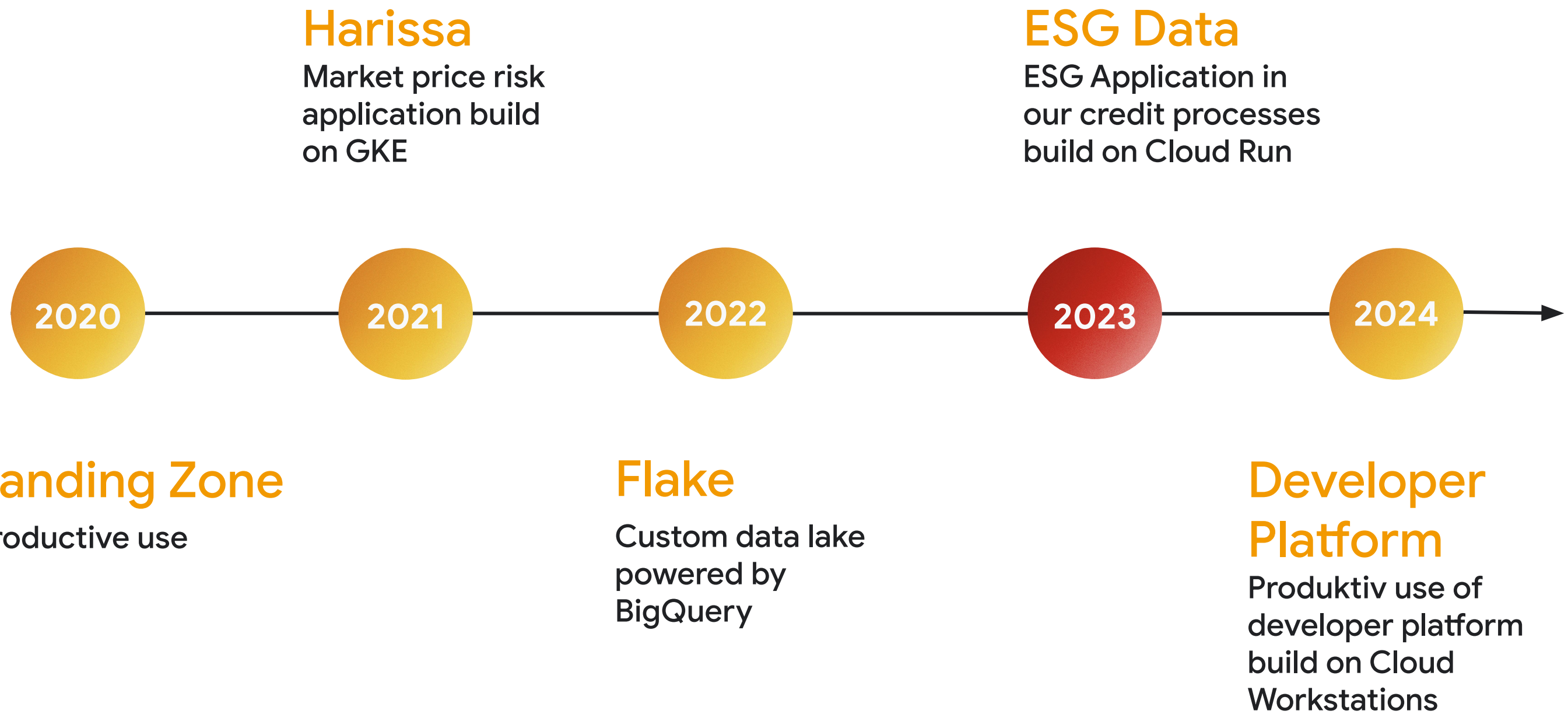






**Industry:** Financial Services  
**Country:** Germany 

 **Customer Award**





# Cloud Run @ DZ BANK

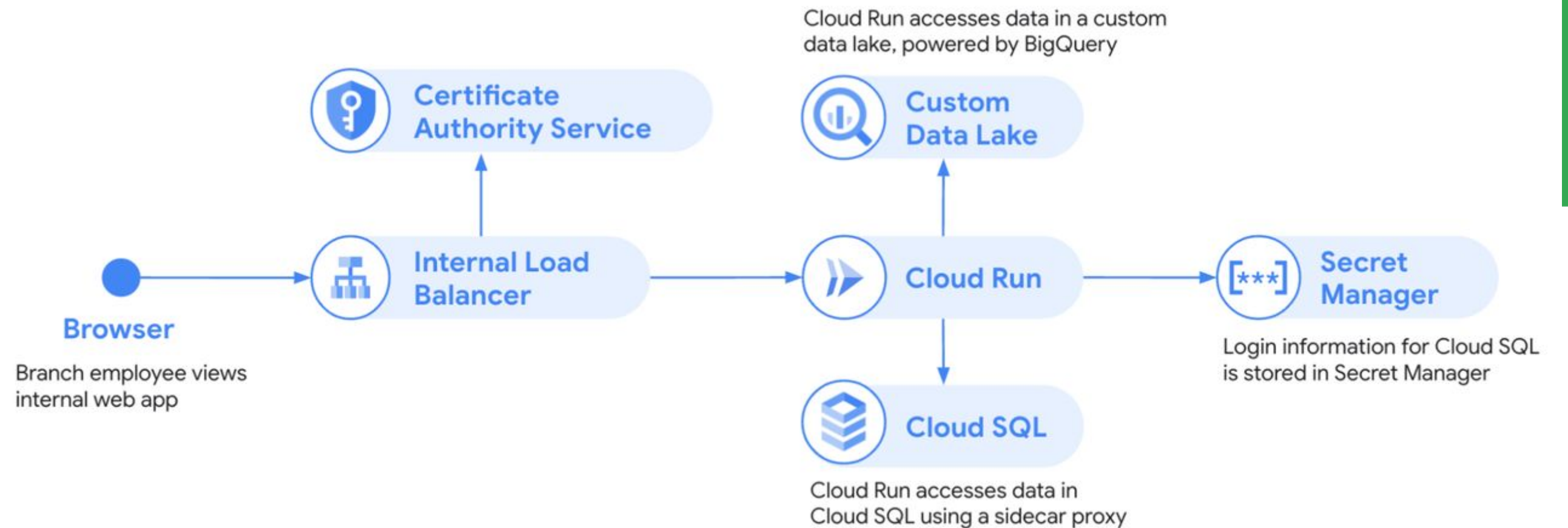
## Use Case: ESG Data

Simple architecture

Met rigorous security and compliance standards

70% toil reduction

90% cost savings





# What we learned

**Safe** Gives us an opportunity to innovate and iterate quickly within regulatory and compliance constraints

**Easy** Automation first with managed services reduces toil and gives our team the gift of time

**Efficient** Serverless mindset enables us cost savings and easy scalability





# In conclusion



## Simplifying App Development

Cloud Run continues to focus on new features to simplify the experience of app developers, including Gemini in Cloud Run and the application canvas

## Enterprise Ready

With built in security and easy integration with existing systems, Cloud Run can reduce toil and cost of demanding enterprise Workloads



**We are interested in  
your feedback!**

**Connect with a  
Product Manager or  
UX researcher.**





# Ready to build what's next?

Tap into **special offers** designed to help you **implement what you learned** at Google Cloud Next.

**Scan the code** to receive personalized guidance from one of our experts.



Or visit [g.co/next/24offers](https://g.co/next/24offers)



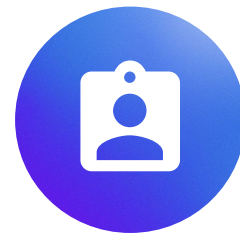
# Continue your learning journey!



## Sessions

**DEV310** – Serverless security like a pro

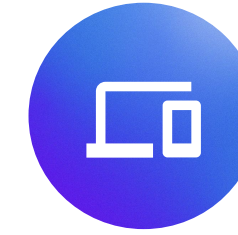
**ARC202** – Design secure enterprise networks for a multi-cloud world



## Sessions

**SEC211** – Next-generation permissions management: Control identities and privileged access to reduce risk

See more from Accenture at **Booth 401**



## Demo in Showcase

**AIPD-110** Enhance consumer experiences with generative AI



Thank you

