Google Cloud

Dataflow for Real-Time ML & Generative AI

Dataflow architecture that ensures low latency predictions even when the decision cycles of ML & GenAI models are long

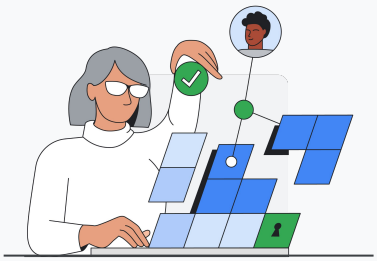
Real-time predictions

Incorporate predictions based on real-time data that enable differentiated product experiences

Instantaneous personalization

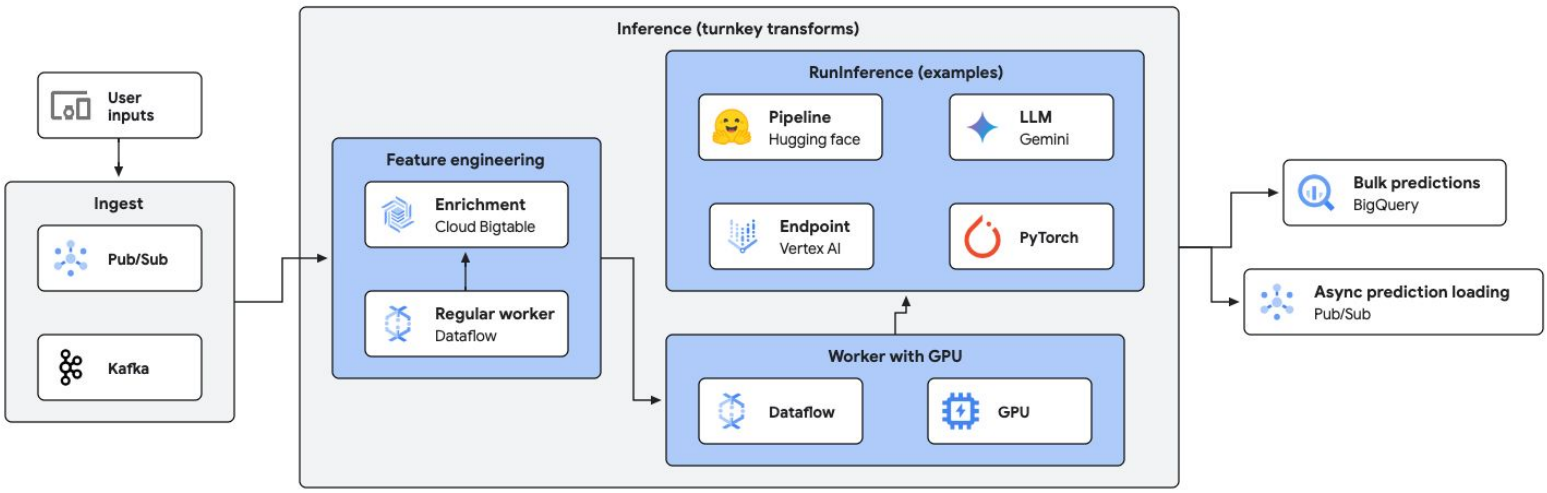
Deliver personalized interactions to improve customer satisfaction without waiting for days for computations

Improved customer support



Predict customer churn risk in real-time and extend offers & targeted measures to decrease churn

Ingest streaming data into Pub/Sub or Kafka. Dataflow can pick up this data, enrich events with static data, then plug into a model of your choice (Vertex AI, Gemini, PyTorch, and Hugging Face) to make predictions. Writing predictions out to Pub/Sub decouples the application from the data processing layer.



Dataflow can deliver on the ML & gen AI potential in Real-Time

1

Retail & E-Commerce

Use Cases

- Real-time personalized recommendations
- Dynamic pricing responding in real-time to fluctuating demand & inventory levels

Value

- Improve customer satisfaction and increase retention rates
- Maximize revenue per user

2

Finance

Use Cases

- Analyze transaction patterns, user behavior, and other data points
- Agents powered by generative AI based on individual user data and goals

Value

- Identify & flag fraudulent activities
- Personalized financial reports, investment recommendations, and retirement planning scenarios

3

Media & entertainment

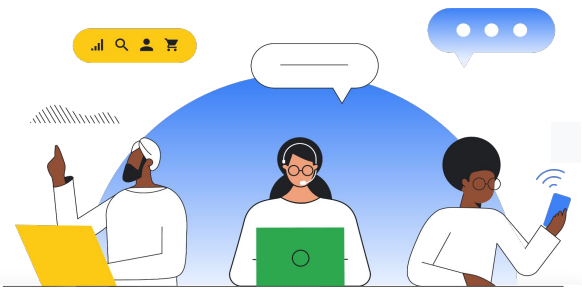
Use Cases

- Integrate generative AI capabilities into product experience
- Analyse real-time data on user behavior & viewing patterns

Value

- Help creators enhance user-generated content
- Make personalized content recommendations

Why Dataflow



Innovate with Google's Dataflow unified stream and batch data processing

Build and scale fast with our open-source compatibility

Help accelerate time-to-production with enterprise-ready streaming

Technical benefits

- **Developer ease of use with turnkey transforms and Notebooks integration**
- **Advanced stream processing:** state & timer APIs, side inputs, connectors ...
- **Cost efficiency:** Right-fitting, GPU support
- **Open-source: support for running inference with Gemma and strong integration with Tensorflow Extended**

Spotify applies heavyweight ML models to process uploaded podcasts and offer relevant snippets for previews. With Dataflow, they have reduced processing time from 2 hours down to 2 minutes.

Go-Jek uses Dataflow for its entire machine learning lifecycle (feature creation, standardization, and inference) to forecast demand and dynamically price services for its on-demand gig platform

HSBC built a new scenario-risk modeling tool with Dataflow that sped up computations by 16x and empowers traders to better manage their portfolios on an intraday basis

Learn More

Let's get started

- **Align on goals** for developer efficiency and key use cases
- Review reference architecture and **implementation checklist**
- Engage with **Google Cloud Consulting** or certified **Google Cloud Partner**
- Activate Google Cloud Consulting service packages to **streamline implementation**

Learn more today