Google Developer Group
Yogyakarta

Reliable and Scalable AI Infrastructure with Event-Driven Architecture on Google Cloud

Ananda Dwi Rahmawati

Google Developer Expert - Cloud

Build with AI

Google Developer Group
Yogyakarta

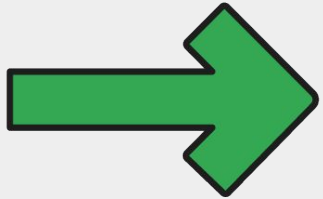
Ananda Dwi Rahmawati

- Cloud & DevOps Engineer, Singapore
- Google Developer Expert Cloud - Modern Architecture
- Master of Computer Science - University of Texas at Austin
- https://linktr.ee/misskecupbung

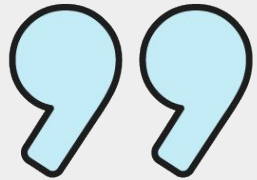
Build with AI

Chapter One

Challenges in AI Infrastructure

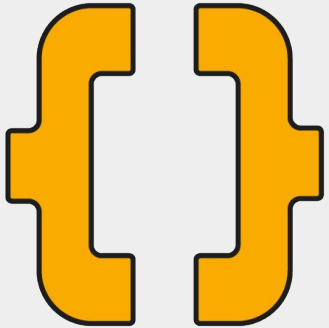


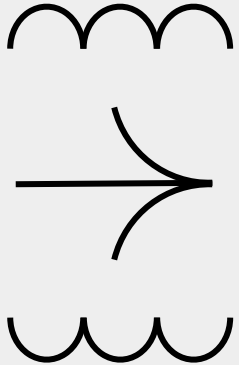
“Building AI infrastructure isn’t just about more compute — it’s about orchestrating complexity, uncertainty, and scale into something reliable and repeatable.”



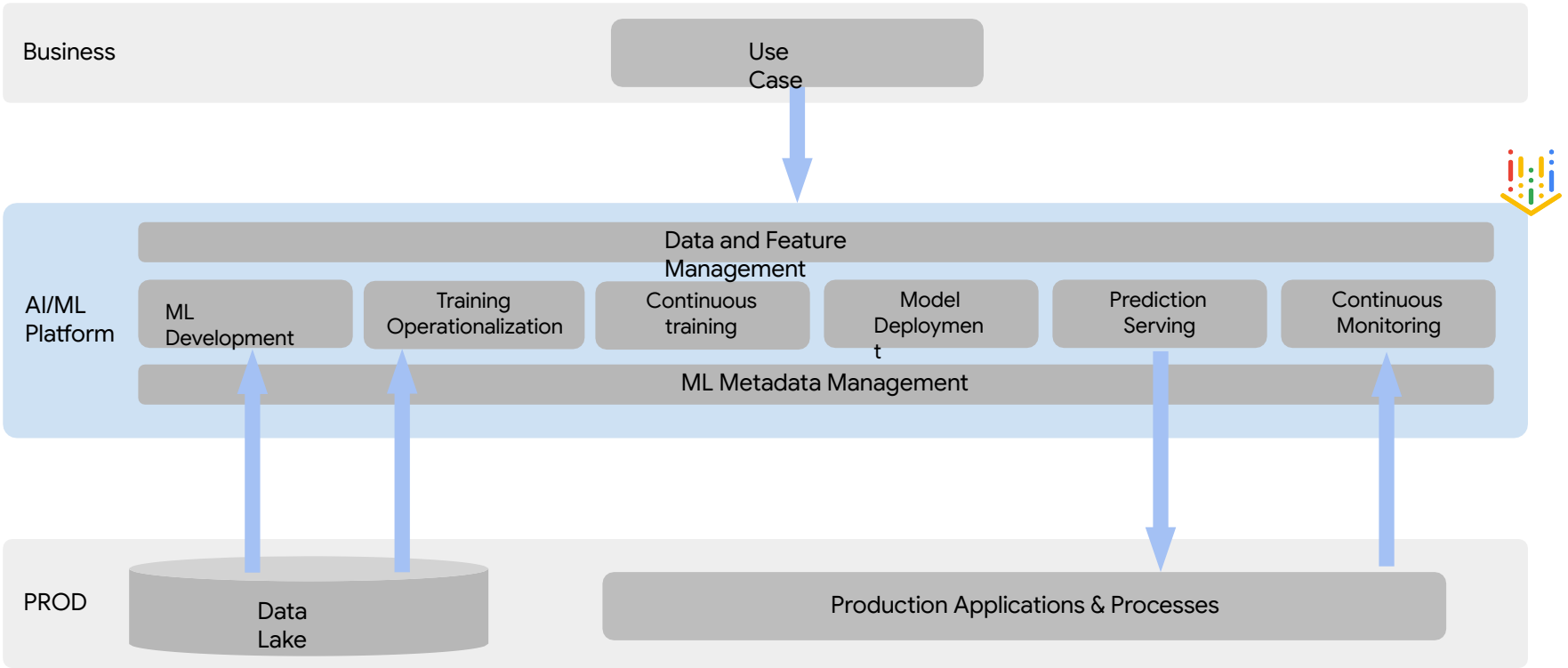
Challenges in AI Infrastructure

- **Complex data workflows**: preprocessing, training, serving.
- Handling large data volumes and **real-time** events.
- **Reliability**: retries, failures, monitoring.
- **Scalability**: burst loads, large model serving.



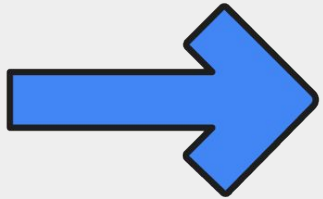


MLOps: quick recap

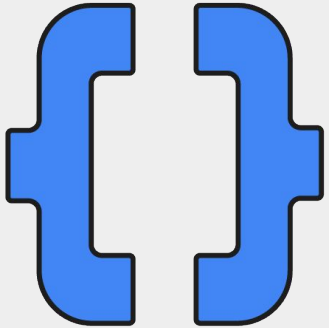


Chapter Two

Event-Driven Architecture (EDA)

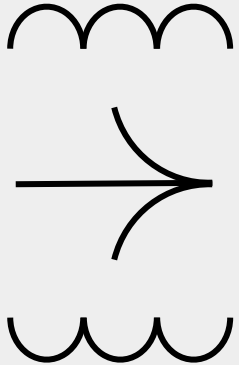


Event-Driven Architecture Overview



What is Event-Driven Architecture?

- Software design where microservices react to **events** (state changes or notifications).
- Events trigger services without services knowing about each other (only event format matters).
- Microservices apply different logic and emit their own events.

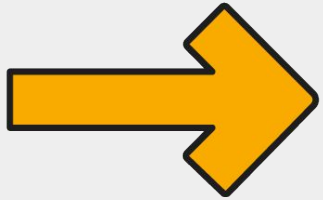


Key Characteristics of Events:

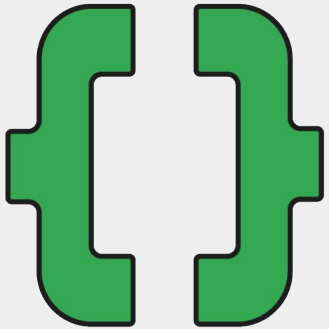
- Record of an immutable fact.
- Persistable indefinitely and re-consumable.
- Occur independently of service logic.

Chapter Three

Why Event-Driven Architecture?

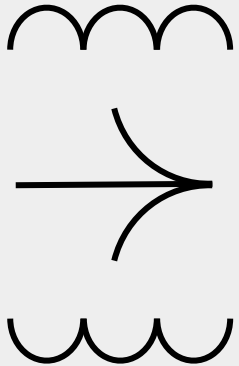


Benefits & Key Considerations



Benefits:

- Loose Coupling: Independent scaling, updating, and deployment.
- Asynchronous & Resilient: Services fail independently; events can be replayed.
- Push-Based, Real-Time: Reduces network load and cost.
- Audit & Event Sourcing: Immutable logs for traceability and state recreation.

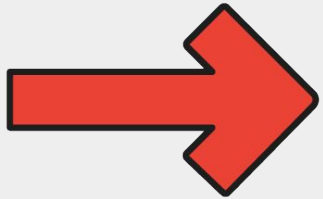


Key Considerations:

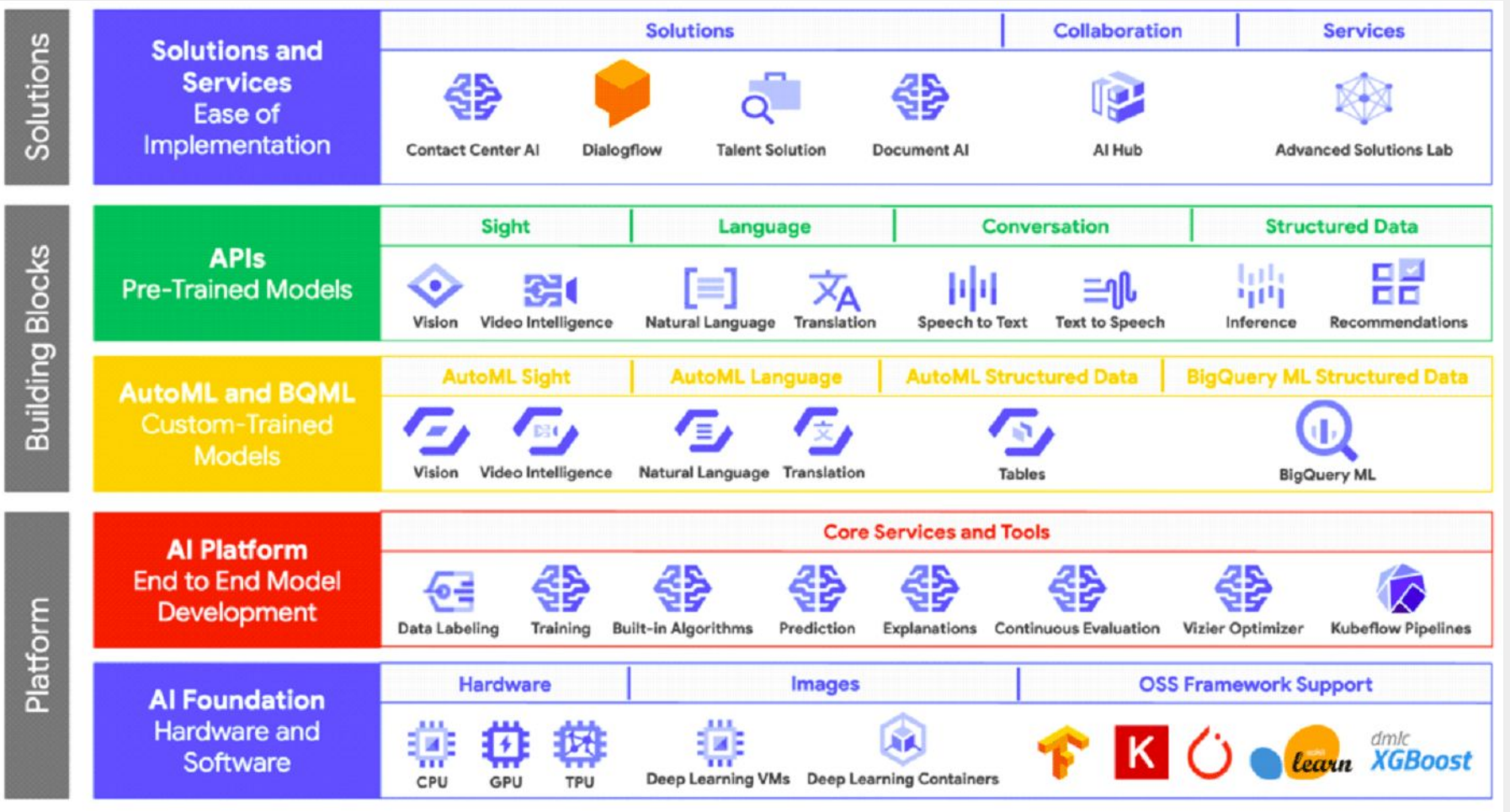
- Ensure reliable event delivery for critical processes.
- Design for asynchronous, scalable handling of requests.
- Implement dynamic monitoring for event flow tracking.
- Plan deduplication and ordering for accurate state rebuilding.

Chapter Four

Google Cloud for Event-Driven AI Infrastructure



Are you ready?

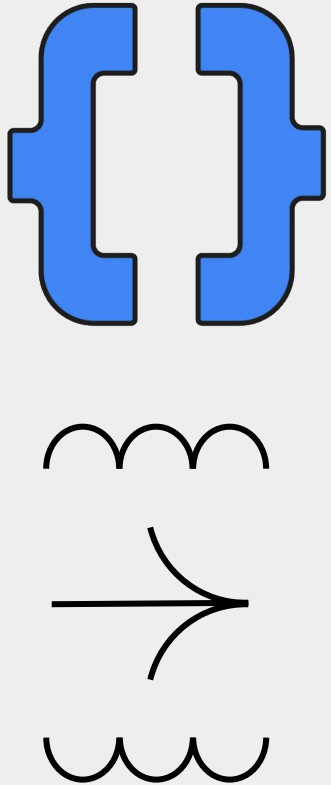


Source: https://www.softmaxai.com/ml-ai-services-by-gcp/

Google Cloud for Event-Driven AI Infrastructure

Key Google Cloud Services for EDA:

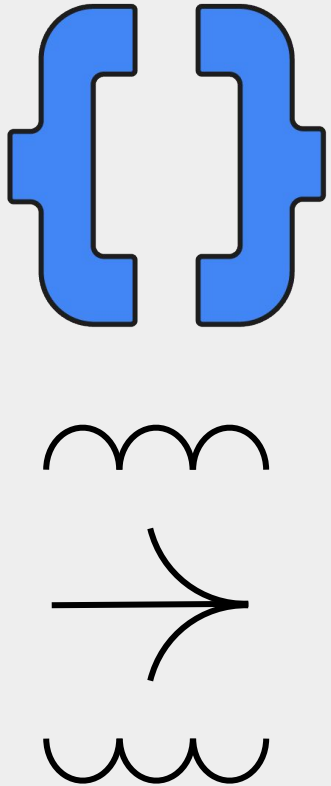
- **Pub/Sub**: Asynchronous messaging service for reliable and scalable data streams.
- **Cloud Functions**: Serverless compute for event-driven logic execution.
- **Cloud Run:** Container-based serverless for more complex event processing.
- **Eventarc**: Globally managed event routing service.
- **Dataflow**: Scalable data processing for event stream analysis.
- **Bigtable/Firestore**: NoSQL databases for scalable state management.
- **Vertex AI**: Google Cloud's unified AI platform and its integration with EDA.



Google Cloud for Event-Driven AI Infrastructure

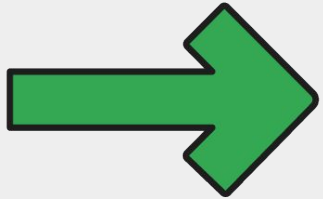
Architectural Patterns for AI with EDA on Google Cloud:

- **Real-time Feature Engineering Pipeline**: Events triggering data transformations and storage.
- **Decoupled Model Training**: Events initiating training jobs based on new data.
- **Asynchronous Inference:** Events triggering model predictions and storing results.
- **Monitoring and Alerting:** Events signaling anomalies and triggering notifications.

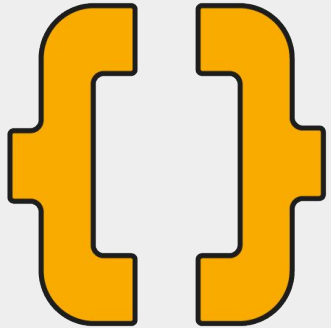


Chapter Five

AI Pipelines with EDA on Google Cloud

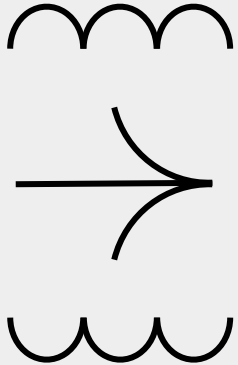


Building Reliable and Scalable AI Pipelines with EDA on Google Cloud



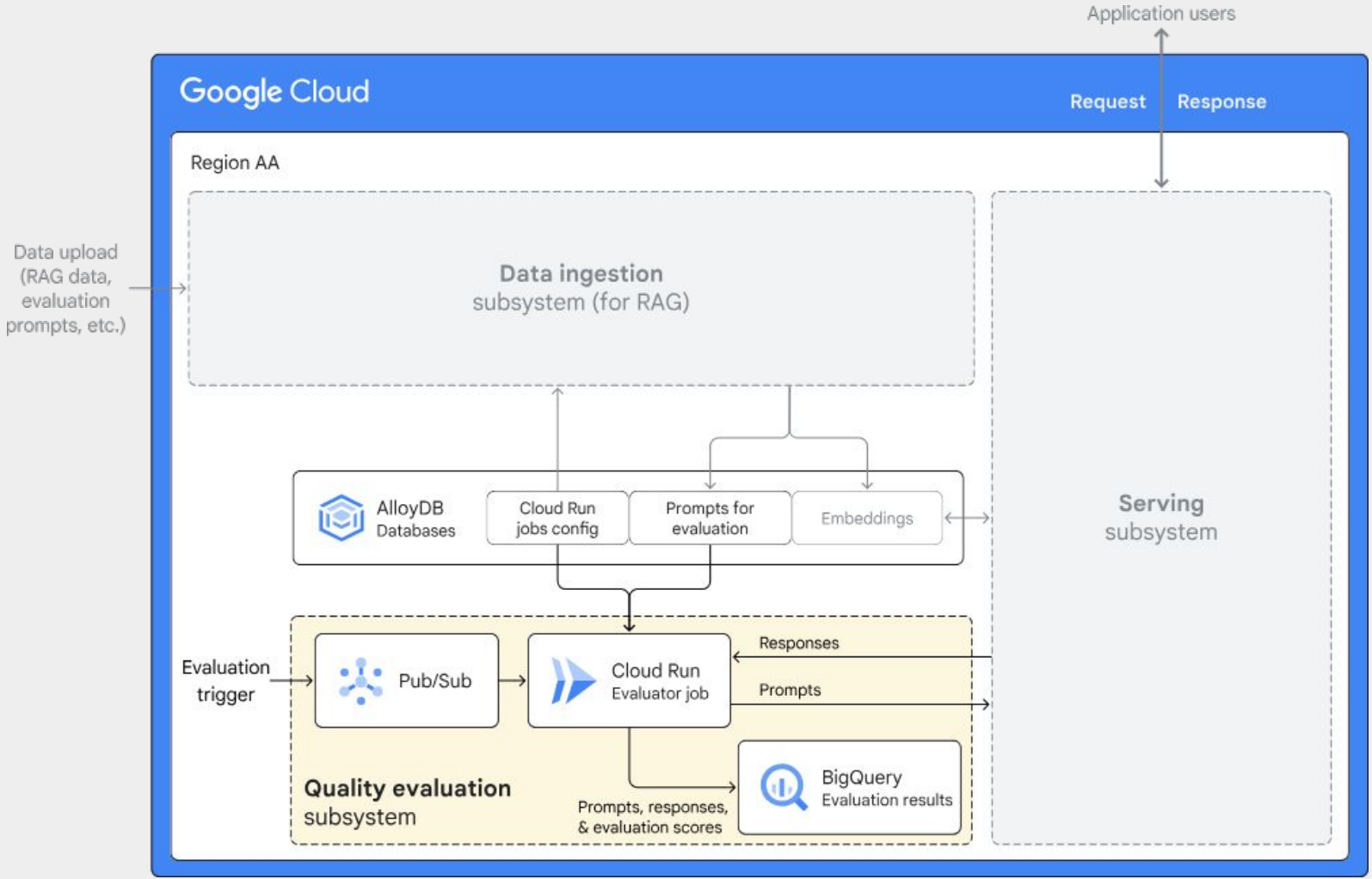
Reliability Considerations:

- Message Durability and Delivery Guarantees (**Pub/Sub**).
- Idempotency in Event Processing (**Cloud Functions/Run**).
- Dead-Letter Queues (DLQs) for handling failed events.
- Monitoring and Logging with Cloud Monitoring and Cloud Logging.
- Error Handling and Retry Mechanisms.

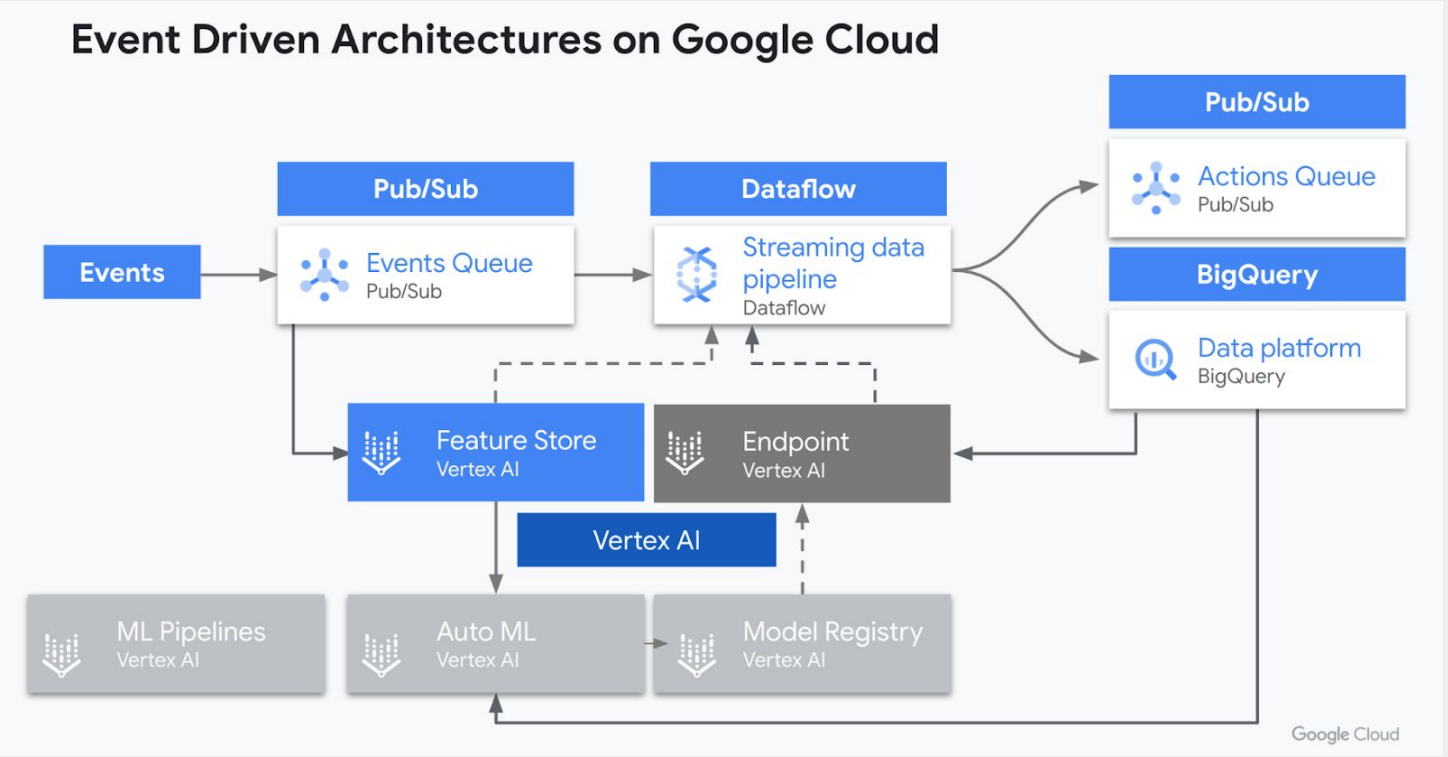


Scalability Considerations:

- Horizontal Scaling of Consumers (**Cloud Functions/Run**).
- Autoscaling of Dataflow jobs.
- Scalable Storage Solutions (**Bigtable/Firestore**).
- Optimizing Event Processing Logic for Performance.

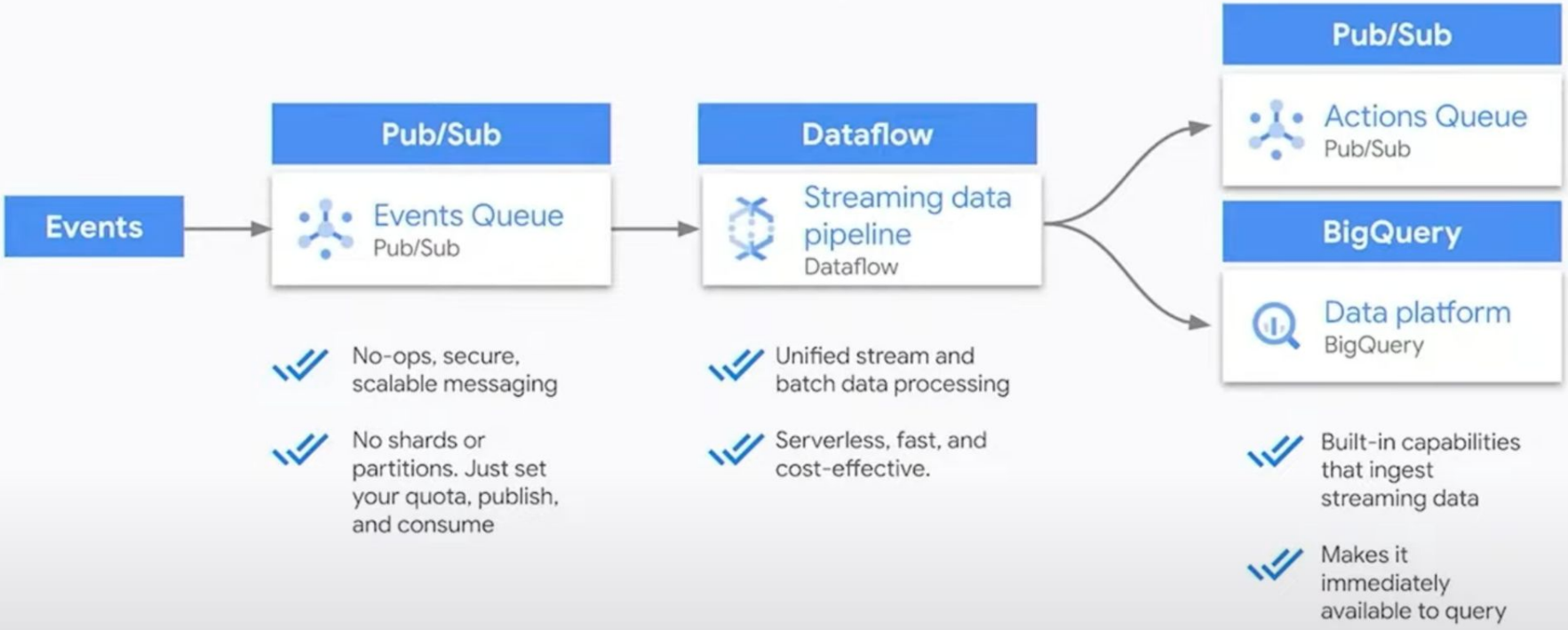

Application users
Google Cloud
Request
Response
Region AA
Data upload (RAG data, evaluation prompts, etc.)
Data ingestion subsystem (for RAG)
AlloyDB Databases
Cloud Run jobs config
Prompts for evaluation
Embeddings
Serving subsystem
Evaluation trigger
Pub/Sub
Cloud Run Evaluator job
Responses
Prompts
Quality evaluation subsystem
Prompts, responses, & evaluation scores
BigQuery Evaluation results

Event-Driven Architecture Overview



Source: https://cloud.google.com/eventarc/docs/event-driven-architectures

Event Driven Architectures on Google Cloud



Thank you!