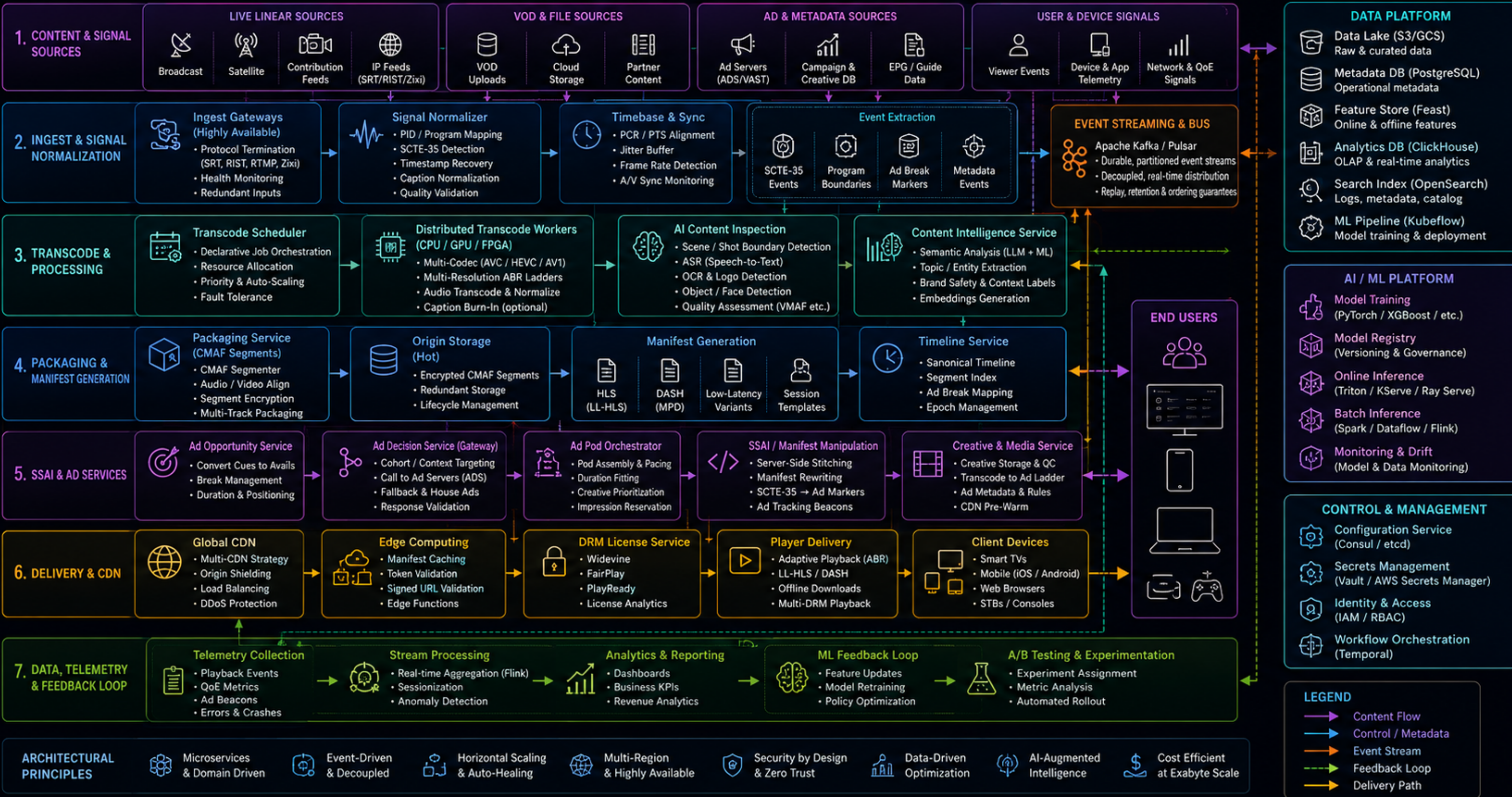


HIGH-LEVEL DISTRIBUTED ARCHITECTURE

AI-NATIVE VIDEO & AD PLATFORM FOR TARGETED SSAI AT SCALE



VISUAL PRODUCTION WORKFLOW AT SCALE

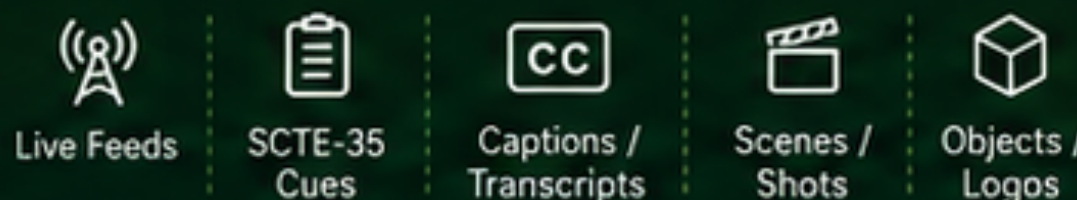
AI Agents + ML Models + Real-time Data + Deterministic Systems = Targeted SSAI at Scale

1. INPUT SIGNALS

VIEWER SIGNALS



CONTENT SIGNALS



ADVERTISING SIGNALS



2. EVENT & STREAMING LAYER

kafka
(or Pulsar)



Durable, scalable
log of all events
(security, replay
& audit)

3. REAL-TIME STREAM PROCESSING & STATE

Apache
Flink



Stateful stream
processing with
exactly-once
semantics

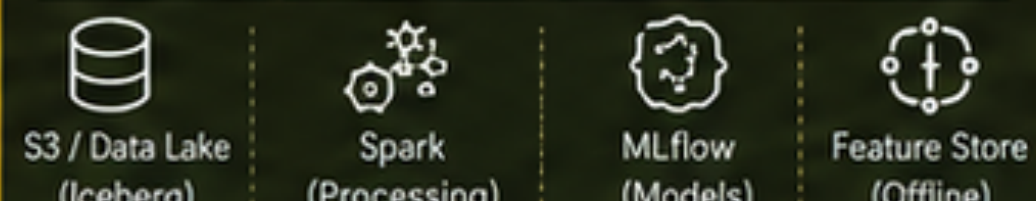
4. FEATURE & DATA PLATFORM

ONLINE FEATURE STORE (Low Latency)



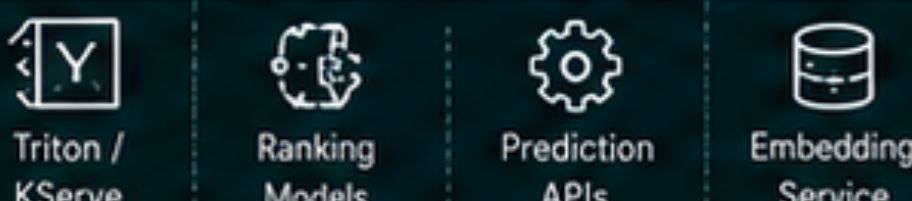
~1-10 ms

OFFLINE DATA PLATFORM (Analytics & ML)



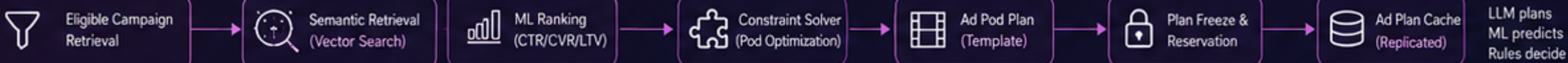
Minutes - Hours

MODEL SERVING LAYER



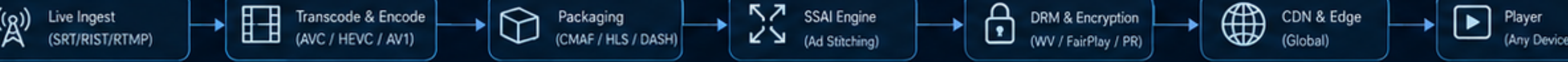
~5-20 ms

5. AI-POWERED DECISIONING PLANE



LLM plans
ML predicts
Rules decide

6. VIDEO DELIVERY PLANE (SSAI)



7. MEASUREMENT & FEEDBACK LOOP



AI AGENTS (CONTROL PLANE)



Content Intelligence Agent

- Understand live content in real-time
- ASR, OCR, CV, LLM semantic analysis
- Output: topics, entities, context



Audience Intelligence Agent

- Build privacy-safe interest profile
- Predict completion & churn risk
- Manage frequency & recency



Campaign Agent

- Translate goals to constraints
- Forecast delivery & pacing
- Optimize bid & budget strategy



Creative Intelligence Agent

- Validate technical quality
- Extract semantics & brand safety
- Normalize & transcode creatives



Ad Pod Planning Agent

- Select best ads for each opportunity
- Optimize revenue & viewer experience
- Propose pod plan



Quality & Remediation Agent

- Detect issues (content/ads/playback)
- Correlate errors & root cause
- Auto-remediate within guardrails



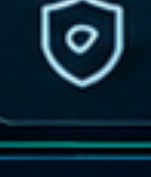
Experimentation Agent

- Design & run experiments
- Measure lift & statistical significance
- Manage treatments & rollouts



Incident Response Agent

- Monitor anomalies & SLOs
- Triage incidents & page ops
- Orchestrate recovery workflows



Agent Orchestrator & Policy Engine

Co-ordinates agents, enforces policies, approvals, audit, and guardrails

KEY PRINCIPLES

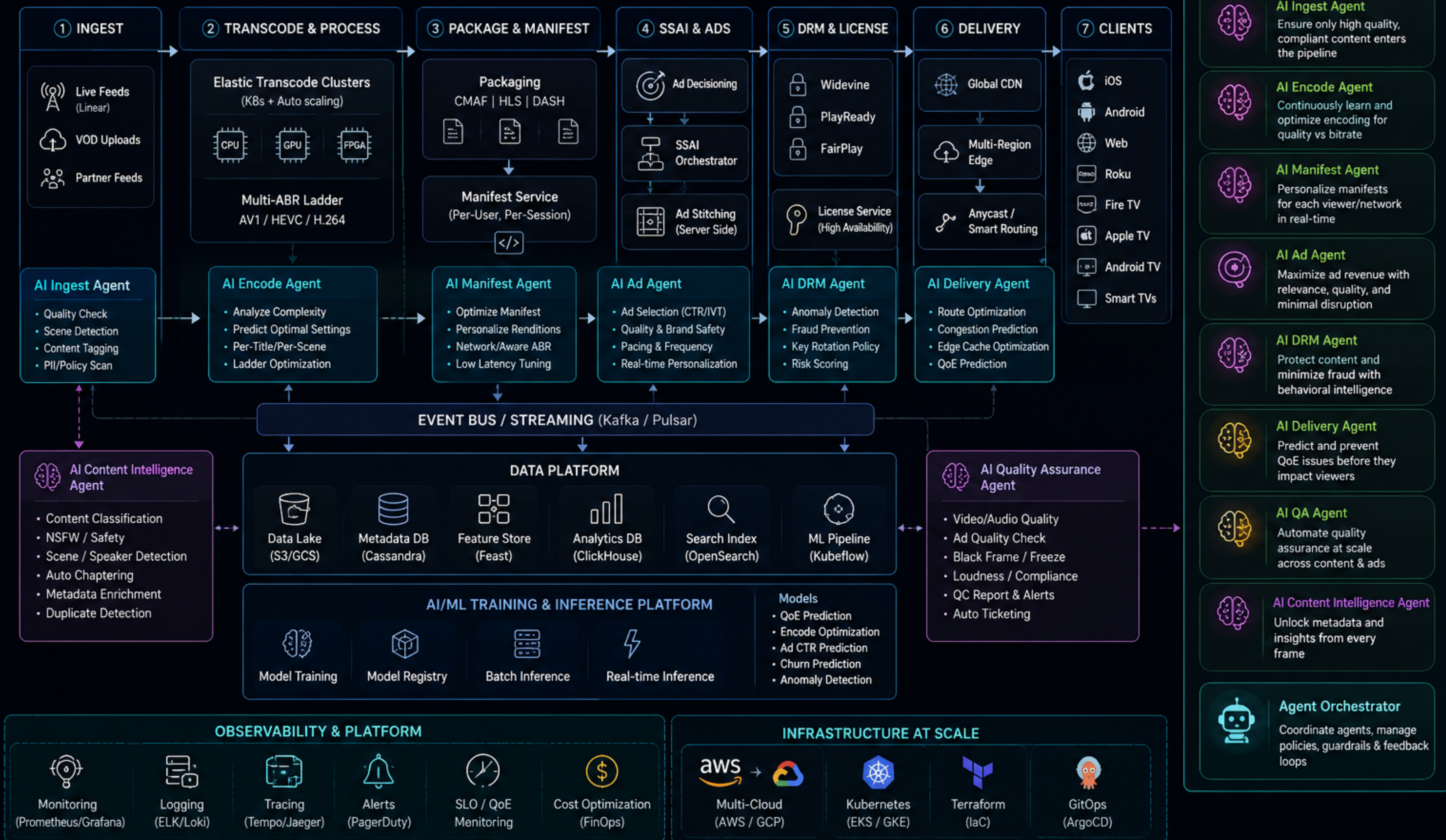
- AI Everywhere
- Real-time Decisions
- Data-Driven Optimization
- Cost Efficient at Exabyte Scale
- High Availability & Resilience
- Security, Privacy & Compliance

FOUNDATION: CLOUD INFRASTRUCTURE & SECURITY



AI-NATIVE VIDEO ENGINEERING PLATFORM @ SCALE

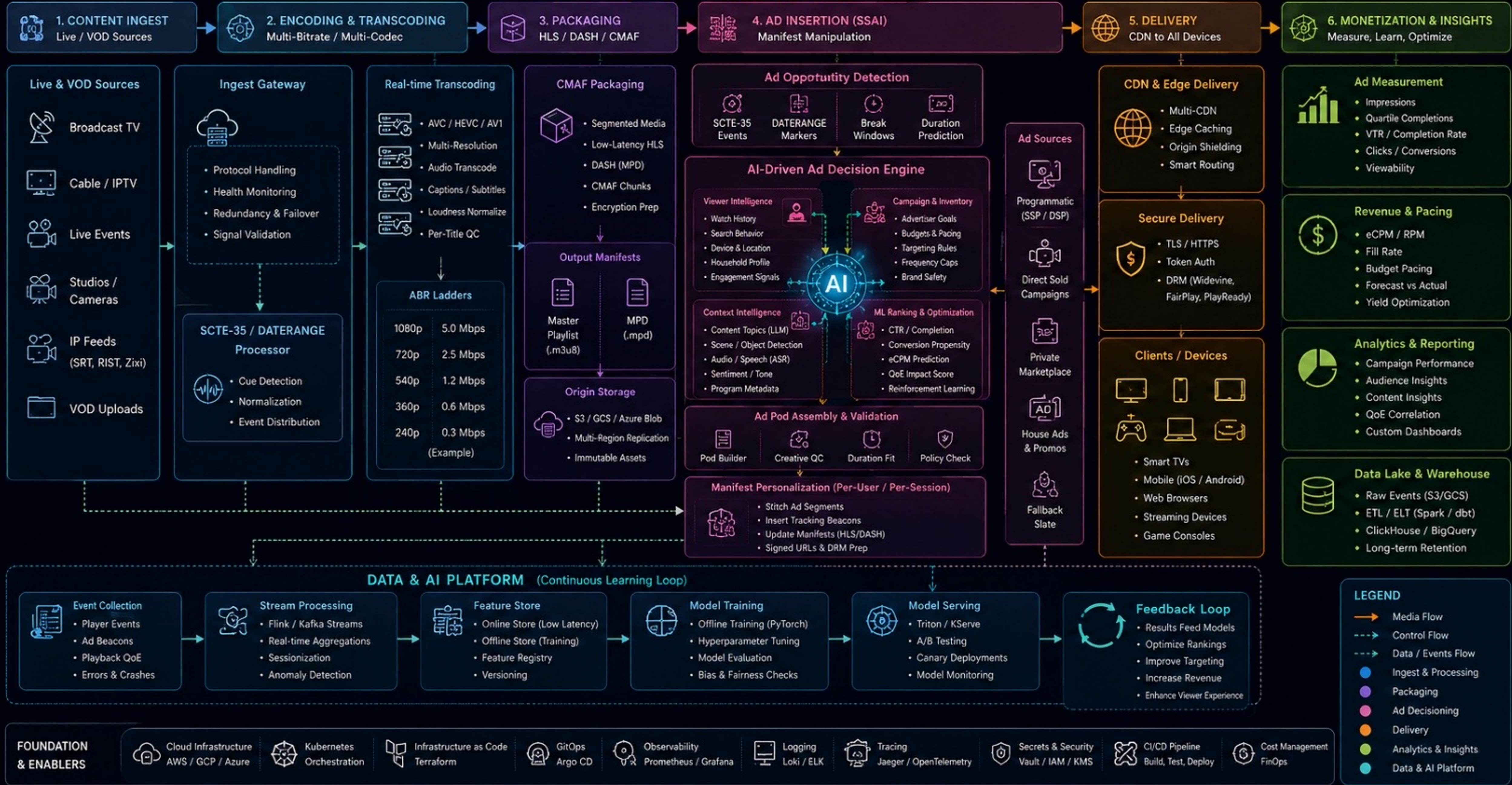
Intelligent Ingest → Transcode → Package → Deliver → Measure → Optimize



KEY PRINCIPLES: 1. AI Everywhere 2. Real-time Decisions 3. Data-Driven Optimization 4. Cost Efficient at Exabyte Scale 5. High Availability & Resilience 6. Security & Compliance by Design

AI-POWERED ADVERTISING PIPELINE

From Live Ingest to Personalized Ads at Scale



RECOMMENDED SERVICE DECOMPOSITION

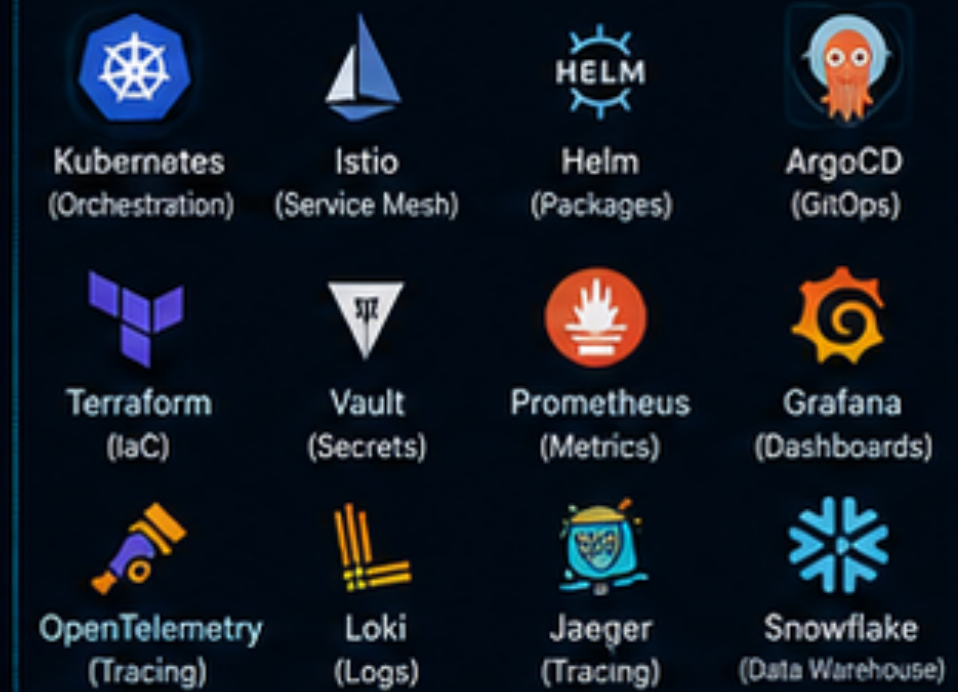
AI-Native Video & Advertisement Platform (Targeted SSAI at Scale)



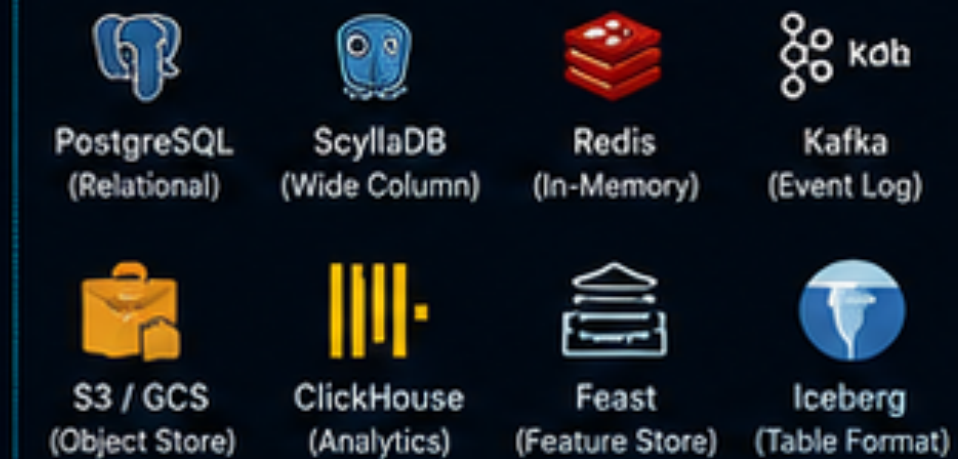
INTER-SERVICE COMMUNICATION

- gRPC (Internal)**
High-performance, strongly typed
- REST/JSON (External)**
Public & partner facing APIs
- Async Events (Kafka)**
Event-driven, decoupled systems
- Schema Registry**
Versioned, backward compatible

SHARED PLATFORMS



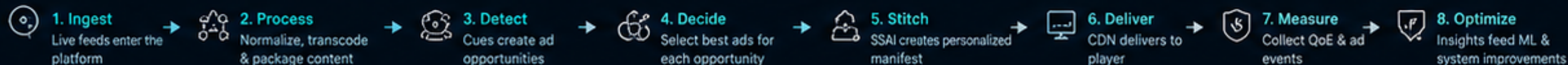
DATA STORES USED



KEY BENEFITS

- ✓ Independent scaling per service
- ✓ Technology best-fit per domain
- ✓ High availability & fault isolation
- ✓ Faster development & deployment
- ✓ Cost efficiency at massive scale

END-TO-END FLOW



MULTI-LAYER TOOL STACK

End-to-End AI-Native Video Engineering Platform for Targeted SSAI at Scale



CROSS-CUTTING CAPABILITIES

 **API & SERVICE LAYER**
REST / gRPC APIs, GraphQL Gateway, BFF, Rate Limiting
Kong / Apigee / Envoy

 **WORKFLOW & JOBS**
Orchestration of long-running jobs, retries, sagas, approvals
Temporal / Airflow / AWS Step Functions

 **POLICY & GOVERNANCE**
Centralized policies, data classification, retention, access control
Open Policy Agent (OPA)

 **EXPERIMENTATION & FEATURE FLAGS**
A/B testing, canary, feature flags, holdouts
LaunchDarkly / Custom

 **NOTIFICATIONS & COLLABORATION**
Alerts, incident mgmt, runbooks, approvals, escalations
PagerDuty / Slack / Microsoft Teams

 **CI/CD & GITOPS**
Build, test, scan, deploy in a secure, repeatable way
GitHub Actions / Argo CD / Flux

INFRASTRUCTURE & RUNTIME FOUNDATION



RECOMMENDED TECHNOLOGY STACK

AI-Native Video Engineering Platform for Targeted SSAI at Scale



CROSS-CUTTING CAPABILITIES

- HIGH PERFORMANCE**
- Low-latency APIs
 - Horizontal Auto-scaling
 - Global Edge Delivery
 - Cache Optimization

- RELIABILITY & RESILIENCE**
- Multi-AZ / Multi-Region
 - Auto Failover
 - Graceful Degradation
 - Chaos Engineering

- COST OPTIMIZATION**
- Right-sizing
 - Spot Instances
 - Lifecycle Policies
 - Cost Monitoring

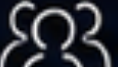
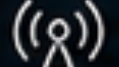
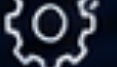
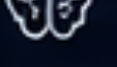
- DATA PRIVACY**
- PII Protection
 - Consent Management
 - Data Minimization
 - Encryption at Rest/In Transit

- DEVEX & AUTOMATION**
- CI/CD (GitHub Actions)
 - Infrastructure as Code
 - Automated Testing
 - Blue/Green Deployments

- OBSERVABILITY CULTURE**
- SLIs / SLOs
 - Error Budgets
 - Dashboards
 - Postmortems

VIDEO STREAMING & ADVERTISING REFERENCE ARCHITECTURE COMPARISON

How Leading Platforms Build, Deliver & Monetize Video at Global Scale

	NETFLIX Netflix	YOUTUBE YouTube	META Meta	Google Google	aws AWS	ARCHITECTURE LAYERS (REFERENCE VIEW)
FOCUS & BUSINESS MODEL	 Subscription (SVOD) Ad-supported (AVOD with Ads tier)	 Ad-supported (AVOD) YouTube Premium (SVOD)	 Ad-supported across Apps (Facebook, Instagram, Reels)	 Ad-supported across Search, YouTube, Display, Discover	 Enable others (Cloud Services) Infrastructure & Managed Services	 Content Sources ↓
CONTENT SOURCES	 Studios, Production Partners, Licensed Content	 UGC, Creators, Partners, Live, Movies & Shows	 UGC, Creators, Live, Short-form (Reels), Partner Content	 UGC, Creators, Live, News, Partners, Movies & Shows	 Broadcasters, Enterprises, OTT, Developers	 Ingest & Capture ↓
INGEST & PROCESSING	 Media Ingest Encoding (FFmpeg), QC ABR Ladder Generation	 Ingestion (Resumable Upload) Transcoding (VP9/AV1/HEVC) Video Enhancement	 Ingest (Live & VOD) Transcoding, Optimization Auto-Caption, Thumbnails	 Global Ingest Transcoding (AV1, VP9, HEVC) Live Processing	 AWS Elemental (MediaConnect, MediaLive, MediaConvert) Automated Workflows	 Transcoding & Processing ↓
STORAGE & MANAGEMENT	 Open Connect CDN (Edge) S3-Compatible Object Store Metadata & Catalog	 Colossus (Distributed Storage) Metadata (Spanner) Global Replication	 TAO Storage (S3-like) Metadata (RocksDB, TAO Metadata) Global Distribution	 Colossus (Distributed Storage) Spanner (Global DB) Bigtable	 Amazon S3 / S3 Glacier DynamoDB / Aurora AWS Catalog (Glue)	 Storage & Metadata ↓
DELIVERY & PLAYBACK	 Open Connect CDN ABR (HLS, DASH) Per-Title Optimization	 Google Global CDN ABR (HLS, DASH) Low-Latency (LL-HLS)	 Edge Network (Global CDN) ABR, Low-Latency Streaming Adaptive Bitrate	 Google Global CDN ABR (HLS, DASH, CMAF) Low-Latency, Widevine	 Amazon CloudFront ABR (HLS, DASH, CMAF) Edge Functions	 Amazon & CDN ↓
ADVERTISING & MONETIZATION	 Server-Side Ad Insertion (SSAI) Personalization, Targeting Ad Tech Partnerships	 Google Ads Contextual + Behavioral Targeting SSAI & Client-side	 Meta Ads Manager Interest Graph, Pixel, SDK Dynamic Creative	 Google Ads & DV360 Performance Max, Auction AI Bidding & Targeting	 AWS Elemental MediaTailor SSAI, SCTE-35, Ad Decisioning VAST / VMAP, CSAI	 Ads & Monetization ↓
AI / ML CAPABILITIES	 Recommendation Engine Encoding Optimization (Per-Title) Anomaly Detection	 Recommendations Content ID, Copyright Matching AV1/VP9 Optimization	 Ranking (Feed/Content) Ad Relevance, ML Models Video Quality & Safety	 RankBrain, Recommendations Ad Ranking, Smart Bidding Video Understanding (ML)	 AI Services (SageMaker, Rekognition, Transcribe, Comprehend) ML for Media Workflows	 Playback (Clients) ↓
OBSERVABILITY & OPERATIONS	 Atlas Monitoring Chaos Engineering Auto-Scaling & Resilience	 Cloud Monitoring SRE Practices, SLOs Massive Scale Automation	 Monitoring (Scuba) Auto-Scaling, Reliability Distributed Tracing	 Cloud Operations Suite SRE, Borg/Kubernetes Observability at Scale	 CloudWatch, X-Ray Auto Scaling, Well-Architected Multi-AZ / Multi-Region	 Monitoring & Analytics ↓
KEY DIFFERENTIATORS	 - Content curation & personalization - Open Connect CDN advantage - High quality Streaming experience	 - Massive UGC scale - Strong discovery & search - Dominant ad platform	 - Social graph targeting - Short-form & Reels scale - Cross-app ad ecosystem	 - Data scale across properties - AI-powered ad bidding - Best-in-class infra & tooling	 - Best-of-breed media services - Flexibility & pay-as-you-go - Enterprise-grade reliability	 AI/ML & Optimization

COMMON ARCHITECTURAL PRINCIPLES ACROSS ALL PLATFORMS

 Global Scale Multi-Region Highly Available	 Microservices Decoupled API-First	 Event-Driven Streaming (Kafka) Real-time Data	 Data-Driven Decisions AI/ML at Every Layer Continuous Learning	 Security & Compliance DRM, Encryption, Access Control	 Observability Metrics, Logs, Traces	 Automation IaC, CI/CD, Self-Healing	 Cost Optimization Right-Sizing, Caching, Tiering
---	---	---	--	---	---	---	--

- Netflix
- YouTube
- Meta
- Google
- AWS

AI-AGENT ARCHITECTURE FOR TARGETED SSAI – 24 PHASES ADOPTION & DELIVERY ROADMAP



A Practical, Safe, and Scalable Path from Foundations to Intelligent, Self-Optimizing Ad Delivery



PHILO GOAL

Deliver the right ad, to the right viewer, in the right context—maximizing revenue without compromising the viewer experience.

GUIDING PRINCIPLES



Viewer First

Protect experience & trust



Privacy by Design

Consent, minimization, pseudonymization



Deterministic Core

Agents recommend, systems decide



Observability by Default

Measure everything, improve continuously



Security & Compliance

Built-in guardrails, audit & governance



Iterate & Learn

Experiment, validate, scale what works

PHASES OVERVIEW



- Foundation (1-6)
- Core Platform (7-12)
- AI Enablement (13-18)
- Optimization (19-24)



STAGE 1: FOUNDATION

Build the strong foundations, data flows, and deterministic core.

PHASES
1-6



1 Assess & Align

- Current state assessment
- Business goals & KPIs
- Stakeholder alignment
- Architecture principles

Deliverable: Target State Vision & Charter



2 Data & Event Strategy

- Event taxonomy
- Data model & schemas
- Data quality standards
- Retention & lifecycle

Deliverable: Data Strategy & Event Catalog



STAGE 2: CORE PLATFORM

Build the ad-tech platform, decisioning services, and operational backbone.

PHASES
7-12



3 Ingestion Backbone

- Deploy Kafka clusters
- Ingest playback, ad, content, & QoE events
- Schema registry

Deliverable: Stable Event Streaming Platform



4 Identity & Consent

- Household identity graph
- Consent management
- Privacy guardrails
- Geo & policy rules

Deliverable: Identity & Consent Service



STAGE 3: AI ENABLEMENT

Introduce AI agents and ML models in assist mode with human-in-the-loop.

PHASES
13-18



5 Storage & Data Lake

- Data lake on S3
- Iceberg/Delta tables
- Metadata management
- Data governance

Deliverable: Governed Data Lake



6 Observability Foundation

- Metrics, logs, traces
- Dashboards & alerts
- Data quality monitors
- SLO definitions

Deliverable: Observability Baseline



STAGE 4: OPTIMIZATION

Scale AI decisions, automation and closed-loop optimization.

PHASES
19-24

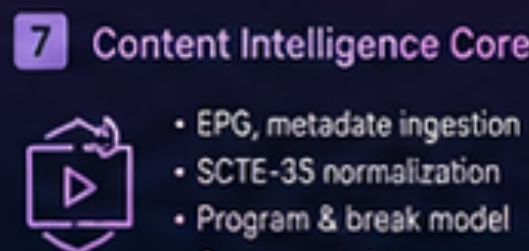
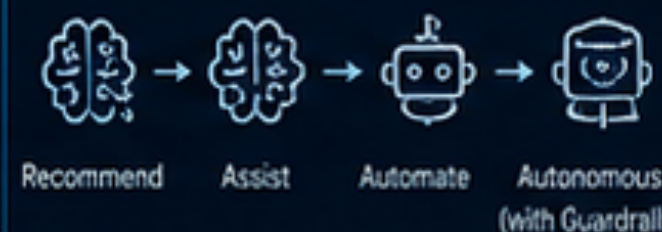
BUSINESS OUTCOMES

- Higher Revenue (↑ eCPM, RPM)
- Better Ad Experience (↑ Completion, ↓ Churn)
- Operational Efficiency (Automation at Scale)
- Smarter Decisions (Data + AI Insights)
- Future-Proof Platform (Scalable & Adaptable)

KEY METRICS IMPACT

- Revenue per Viewer Hour +15% to +35%
- Ad Completion Rate +10% to +25%
- Fill Rate +10% to +20%
- Viewer Churn -5% to -15%
- Operational Cost -10% to -30%

AI AGENT MATURITY PROGRESSION



7 Content Intelligence Core

- EPG, metadata ingestion
- SCTE-3S normalization
- Program & break model
- Content catalog

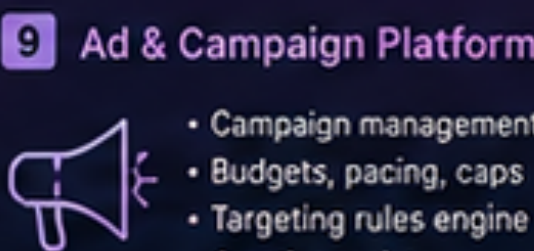
Deliverable: Content Intelligence Service



8 Session & Profile Service

- Session management
- Viewer/household profile
- Real-time state store
- Device & network info

Deliverable: Session & Profile Service



9 Ad & Campaign Platform

- Campaign management
- Budgets, pacing, caps
- Targeting rules engine
- Creative registry

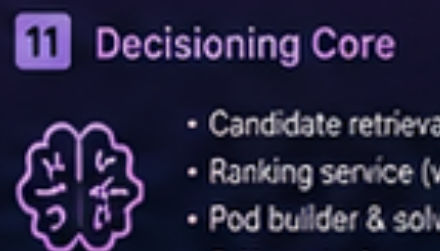
Deliverable: Campaign Management Platform



10 Eligibility & Policy Engine

- Eligibility filtering
- Frequency caps
- Brand safety rules
- Policy enforcement

Deliverable: Policy & Eligibility Service



11 Decisioning Core

- Candidate retrieval
- Ranking service (v1)
- Pod builder & solver
- Fallback logic

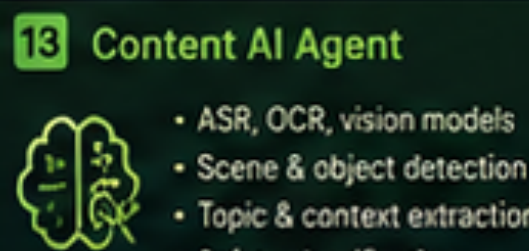
Deliverable: Decisioning Microservices



12 SSAI & Delivery Core

- Manifest stitching (HLS/DASH)
- SSAI ad insertion
- DRM & key service
- CDN integration

Deliverable: SSAI Delivery Platform



13 Content AI Agent

- ASR, OCR, vision models
- Scene & object detection
- Topic & context extraction
- Safety classification

Deliverable: Content AI Agent (Assist Mode)



14 Audience AI Agent

- Interest modeling
- Propensity predictions
- Ad completion & churn
- Privacy-aware features

Deliverable: Audience AI Agent (Assist Mode)



15 Creative AI Agent

- Creative analysis
- Brand & category tagging
- Quality scoring
- Auto QC recommendations

Deliverable: Creative AI Agent (Assist Mode)



16 Campaign Copilot Agent

- Goal to strategy mapping
- Forecasting & pacing
- Inventory simulation
- What-if analysis

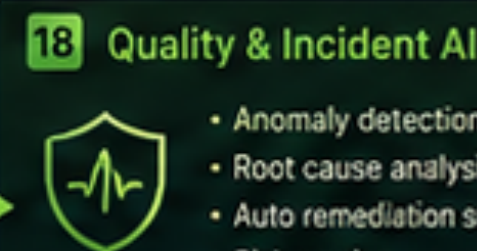
Deliverable: Campaign Copilot (Human-in-loop)



17 Pod Planning AI Agent

- Contextual relevance
- Multi-objective scoring
- Pod optimization
- Reason code generation

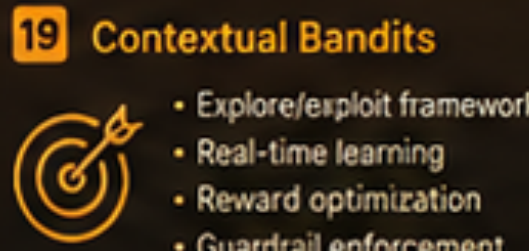
Deliverable: Pod Planning Agent (Assist Mode)



18 Quality & Incident AI Agent

- Anomaly detection
- Root cause analysis
- Auto remediation suggestions
- Risk scoring

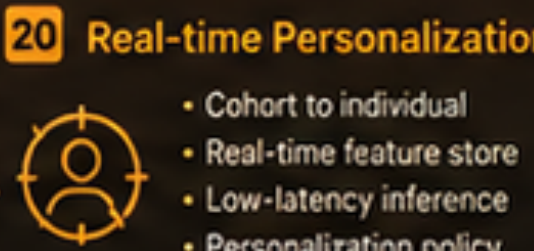
Deliverable: Quality AI Agent (Assist Mode)



19 Contextual Bandits

- Explore/exploit framework
- Real-time learning
- Reward optimization
- Guardrail enforcement

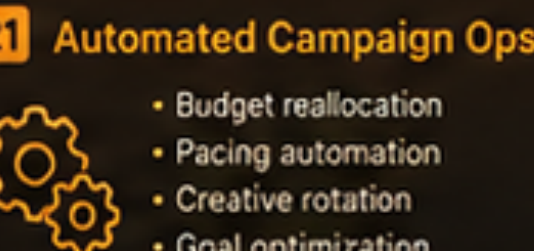
Deliverable: Bandit Optimization (Canary)



20 Real-time Personalization

- Cohort to individual
- Real-time feature store
- Low-latency inference
- Personalization policy

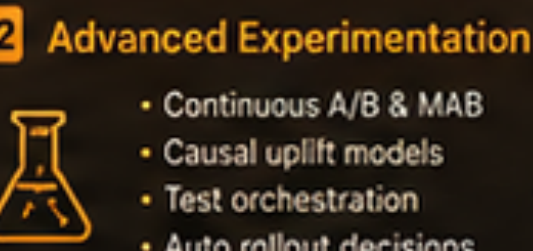
Deliverable: Real-time Personalization



21 Automated Campaign Ops

- Budget reallocation
- Pacing automation
- Creative rotation
- Goal optimization

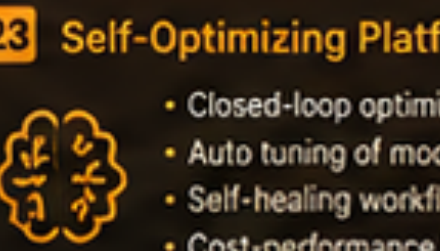
Deliverable: Automated Campaign Operations



22 Advanced Experimentation

- Continuous A/B & MAB
- Causal uplift models
- Test orchestration
- Auto rollout decisions

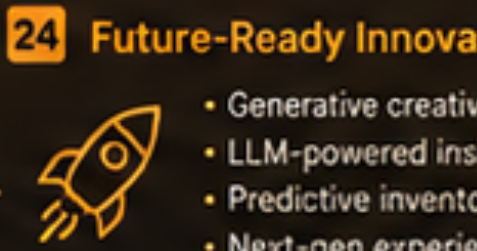
Deliverable: Experimentation Platform 2.0



23 Self-Optimizing Platform

- Closed-loop optimization
- Auto tuning of models
- Self-healing workflows
- Cost-performance balance

Deliverable: Self-Optimizing Platform



24 Future-Ready Innovation

- Generative creative tools
- LLM-powered insights
- Predictive inventory
- Next-gen experiences

Deliverable: Innovation Roadmap & Lab

PLATFORM FOUNDATION (ENABLERS ACROSS ALL PHASES)



Cloud Native on AWS
EKS, ECS, Lambda,
S3, Aurora, DynamoDB



Security & Compliance
IAM, KMS, Vault,
Encryption, Audit



CI/CD & GitOps
CodePipeline, ArgoCD,
Blue/Green Deployments



Data Quality & Governance
Great Expectations,
Data Catalog, Lineage



ML Platform
SageMaker, Feature Store,
Model Registry, Monitoring



Observability
CloudWatch, Prometheus,
Grafana, OpenTelemetry



FinOps
Cost Visibility, Budgets,
Optimization Controls

LEGEND

- Foundation
- Core Platform
- AI Enablement
- Optimization



People / Process



Data / Storage



AI / ML



Streaming / Real-time



Security / Governance



Delivery / Playback

FLOW LEGEND

→ Primary Flow --> Data Flow <--> Feedback Loop