DECISIONFLOW

Technical Architecture Document

MVP system design, non-functional requirements, security and operations

Phase 2 - Plan

Product	DecisionFlow
Document code	DF-P2-04
Version	2.0
Status	Final Public Portfolio Edition — Historical Planning Baseline
Author	Min Thu Kyaw
Date	July 2026

Evidence statement

This public portfolio edition distinguishes actual evidence, assumed continuity steps, simulated case-study outcomes, targets, and recommendations. Simulated customers, commercial results, benchmarks, incidents, quotations, and later-lifecycle outcomes are illustrative and must not be interpreted as results from a live commercial product. Evidence labels are retained throughout the package to support transparent and responsible interpretation.

Document Control

Purpose	Provide implementation guidance for a secure, observable, scalable and model-agnostic MVP architecture.
Primary audience	Engineering, AI/ML, DevOps, QA, security and technical leadership.
Related documents	DF-P2-01, DF-P2-02, DF-P2-05 and DF-P2-07.
Review cadence	Update when scope, evidence, architecture, metrics or major decisions change.
Approval state	Historical phase-gate outcome: Conditional Go to Develop.

Revision History

Version	2.0	Owner	Change
1.0	July 2026	Min Thu Kyaw	Restructured Phase 2 planning package; consolidated and standardized source artifacts.

Contents

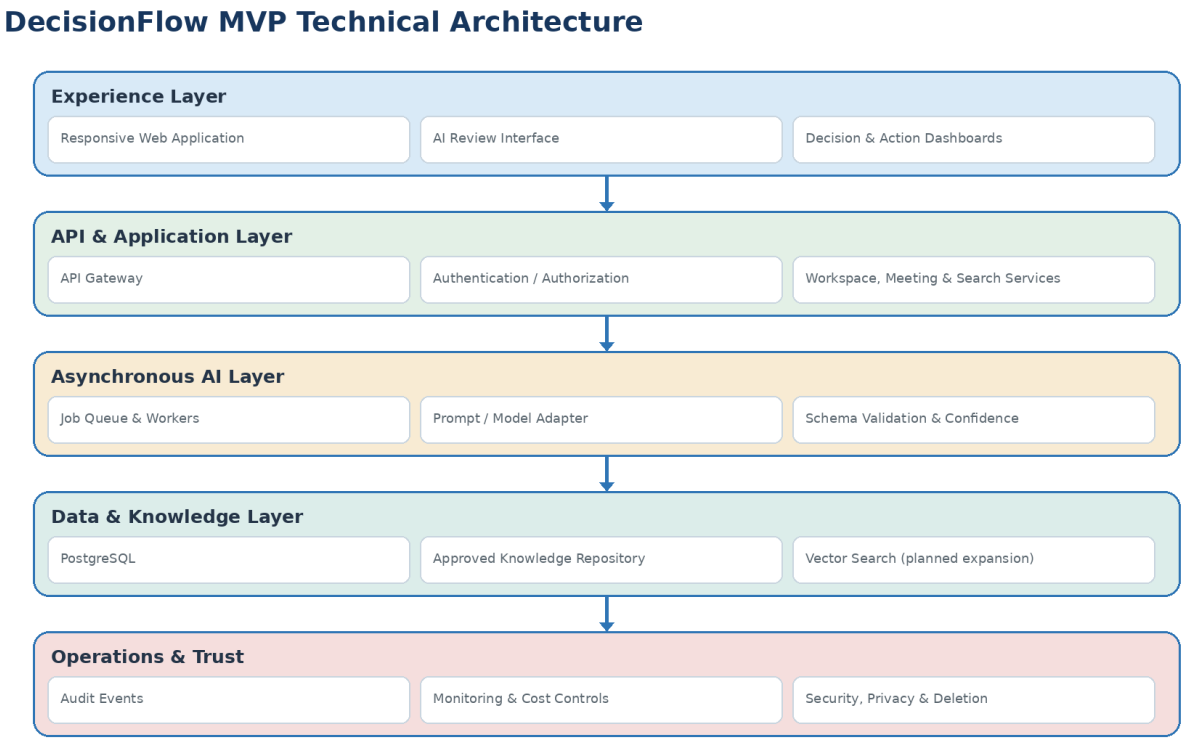


Figure 1. Proposed high-level MVP architecture.

System Architecture and Data Flow

Purpose

This document defines the high-level technical architecture for DecisionFlow, including system components, data flow, AI services, integrations, and deployment strategy. The objective is to provide a shared understanding between Product, Engineering, AI/ML, and DevOps teams before implementation begins.

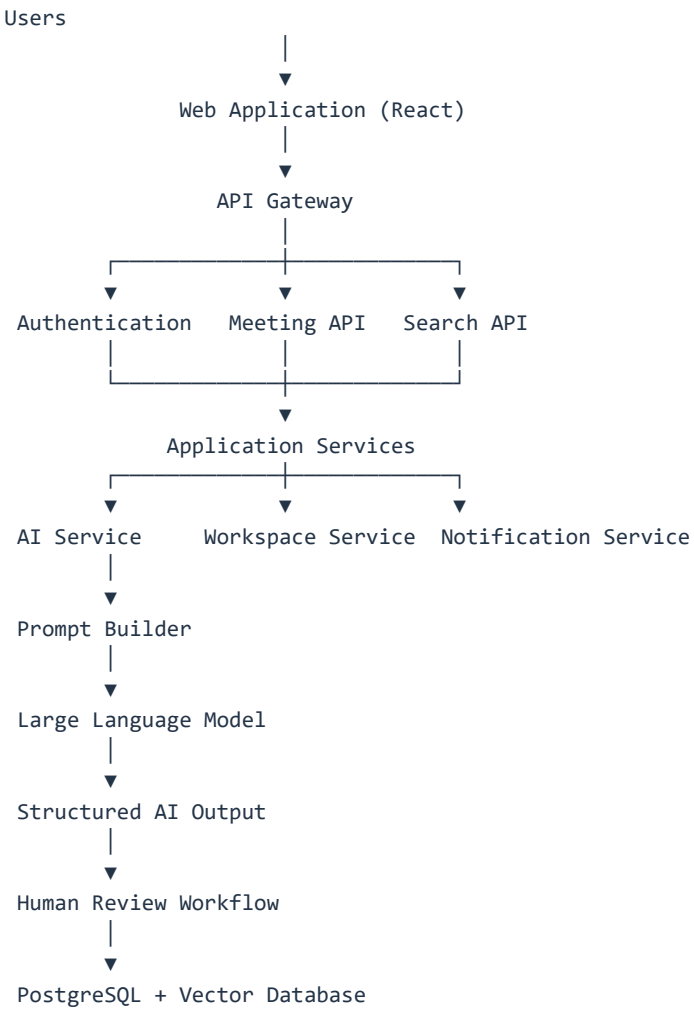
The architecture prioritizes scalability, modularity, reliability, and flexibility while enabling future enhancements without major redesign.

Architecture Principles

DecisionFlow is designed around the following principles:

- API-first architecture
- AI as a modular service
- Cloud-native deployment
- Asynchronous processing for long-running AI tasks
- Vendor-agnostic AI model integration
- Secure-by-design
- Observable and scalable infrastructure

High-Level Architecture



System Components

Web Application

Responsibilities:

- User authentication
- Workspace management
- Transcript upload
- AI review interface
- Decision repository
- Action dashboard

Technology (MVP):

- React
- TypeScript
- Tailwind CSS

Backend API

Responsibilities:

- Business logic

- Authentication
- Authorization
- API orchestration
- Data validation
- Integration with AI services

Technology (MVP):

- FastAPI (Python)

AI Service

Responsibilities:

- Prompt construction
- Model invocation
- Response validation
- Confidence scoring
- Structured JSON generation

The AI service is isolated from business logic to allow independent evolution.

Background Processing

Responsibilities:

- Long-running AI inference
- Retry failed jobs
- Queue management
- Progress tracking

Asynchronous processing ensures that users do not wait for lengthy AI operations.

Database Layer

Relational Database

Stores:

- Users
- Workspaces
- Meetings
- Decisions
- Action items
- Audit logs

Technology:

- PostgreSQL

Vector Database (Future)

Stores:

- Meeting embeddings
- Semantic search vectors
- Organizational knowledge

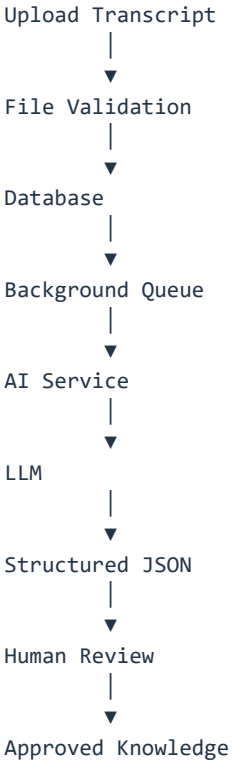
Potential technologies:

- pgvector

- Pinecone
- Weaviate

Data Flow

Meeting Processing Flow



AI Pipeline

Step 1

Transcript preprocessing

- Clean formatting
- Remove noise
- Split into manageable chunks

Step 2

Prompt construction

Includes:

- System prompt
- User prompt
- Meeting metadata
- Output schema

Step 3

LLM inference

The model returns structured JSON containing:

- Decisions
- Action items
- Owners
- Due dates
- Confidence scores

Step 4

Validation

The backend validates:

- JSON schema
- Required fields
- Confidence values
- Duplicate entries

Step 5

Human review

Users approve, edit, or reject AI outputs before they become organizational knowledge.

Security Architecture

Security measures include:

- JWT-based authentication
- Role-based access control (RBAC)
- TLS encryption for data in transit
- Encryption at rest
- Secure secret management
- Audit logging
- Rate limiting for public APIs

API Design

RESTful APIs are organized by domain.

Examples:

Domain	Example Endpoints
Authentication	/auth/login, /auth/logout
Workspaces	/workspaces
Meetings	/meetings
AI Processing	/meetings/{id}/analyze
Decisions	/decisions
Action Items	/actions

External Integrations

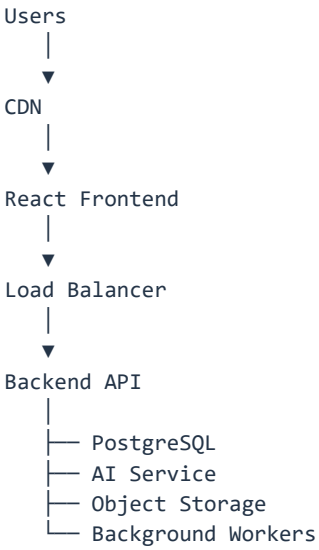
MVP

- LLM provider
- Authentication provider
- Cloud object storage

Future

- Google Calendar
- Microsoft Outlook
- Slack
- Microsoft Teams
- Zoom
- Google Meet
- Jira
- Asana
- Linear
- Notion

Deployment Architecture



Scalability Strategy

The architecture supports independent scaling of:

- Web frontend
- API services
- AI workers
- Search services
- Background processing

This enables efficient resource allocation as product usage grows.

Monitoring & Observability

Key metrics monitored include:

- API latency
- AI processing time
- Error rate
- Queue length
- Infrastructure utilization
- Decision approval rate
- AI confidence distribution

Logs, metrics, and alerts provide operational visibility and support rapid incident response.

Technical Risks

Risk	Mitigation
AI provider outage	Support multiple model providers and retry logic
Long inference times	Background processing and progress indicators
Increasing infrastructure costs	Optimize prompts, caching, and model routing
Large transcript processing	Chunking and asynchronous workflows
Schema inconsistencies	Strict validation before persistence

Architecture Decisions

Decision	Rationale
API-first design	Supports future web, mobile, and third-party integrations
Asynchronous AI processing	Prevents blocking the user interface
Human approval workflow	Improves trust and reduces AI risk
Modular AI service	Enables independent model evolution
Model-agnostic architecture	Reduces vendor lock-in and future migration costs

Success Criteria

The technical architecture is considered successful when:

- Core workflows support MVP requirements.
- AI services can be updated independently.
- System scales without major architectural changes.
- Security and privacy requirements are satisfied.
- Operational monitoring provides actionable insights.
- Future integrations can be added with minimal disruption.

Summary

DecisionFlow's technical architecture is designed to support reliable AI-powered decision intelligence while remaining flexible for future growth. By separating AI services from core business logic, using asynchronous processing, and adopting a modular, cloud-native architecture, the platform balances rapid MVP delivery with long-term scalability, maintainability, and operational resilience.

Non-Functional Requirements

Purpose

This document defines the quality attributes and operational characteristics that DecisionFlow must meet to deliver a reliable, secure, and scalable user experience. Unlike functional requirements, which describe system behavior, non-functional requirements specify performance expectations, operational constraints, and quality standards.

These requirements provide engineering teams with measurable targets for architecture, implementation, testing, and operations.

NFR Categories

DecisionFlow's non-functional requirements are organized into the following categories:

- Performance
- Reliability & Availability
- Scalability
- Security
- Privacy & Compliance
- Usability & Accessibility
- Maintainability
- Observability
- AI-Specific Requirements

Performance Requirements

NFR-001 — Application Response Time

The application shall respond to standard user interactions within 500 milliseconds under normal operating conditions.

Examples

- Opening dashboard
- Searching decisions
- Filtering action items
- Loading meeting details

NFR-002 — AI Processing Time

The AI pipeline shall complete transcript analysis within:

Transcript Length	Target Processing Time
Up to 15 minutes	≤ 15 seconds
15–30 minutes	≤ 30 seconds
30–60 minutes	≤ 60 seconds

NFR-003 — Search Performance

Searching the knowledge repository shall return results within 1 second for datasets containing up to 100,000 approved decisions.

NFR-004 — Concurrent Users

The MVP shall support at least 500 concurrent active users without significant degradation in performance.

Reliability & Availability

NFR-101 — Service Availability

The platform shall maintain 99.9% monthly uptime, excluding scheduled maintenance.

NFR-102 — Fault Recovery

If AI processing fails, the system shall:

- Notify the user.
- Preserve uploaded transcripts.
- Allow retry without requiring another upload.

NFR-103 — Data Integrity

Approved decisions, action items, and audit records shall not be lost due to application failures.

Scalability

NFR-201 — Horizontal Scalability

The application architecture shall support horizontal scaling of:

- API services
- AI processing workers
- Search services
- Background jobs

without requiring significant architectural redesign.

NFR-202 — Storage Growth

The system shall support growth to:

- 1 million meeting transcripts
- 10 million approved decisions
- 50 million action items

through scalable cloud storage.

Security

NFR-301 — Authentication

Only authenticated users shall access workspace data.

NFR-302 — Authorization

Workspace users shall only access resources belonging to their authorized workspace.

NFR-303 — Encryption

The platform shall encrypt:

- Data in transit using TLS.
- Sensitive data at rest using industry-standard encryption.

NFR-304 — Audit Logging

Security-relevant events shall be recorded, including:

- Login attempts
- Workspace administration changes
- Decision approvals
- Permission changes

NFR-305 — Secret Management

API keys, tokens, and credentials shall never be stored in application code or exposed to end users.

Privacy & Compliance

NFR-401 — Data Ownership

Organizations retain ownership of all uploaded meeting data.

NFR-402 — AI Data Usage

Meeting transcripts shall not be used to train foundation models without explicit customer consent.

NFR-403 — Data Deletion

Users shall be able to permanently delete meetings and associated AI outputs.

NFR-404 — Compliance Readiness

The architecture should support future compliance initiatives such as:

- GDPR
- SOC 2
- ISO 27001

without requiring major redesign.

Usability & Accessibility

NFR-501 — Ease of Use

New users shall complete onboarding and upload their first meeting within 15 minutes without external assistance.

NFR-502 — Responsive Design

The web application shall function correctly on common desktop and tablet screen sizes.

NFR-503 — Accessibility

The interface should align with WCAG 2.1 Level AA guidance, including:

- Keyboard navigation
- Sufficient color contrast
- Screen reader compatibility
- Meaningful labels for interactive elements

Maintainability

NFR-601 — Modular Architecture

The system shall separate:

- Frontend
- Backend
- AI services
- Background workers
- Data storage

to improve maintainability and independent deployment.

NFR-602 — API Design

All APIs shall follow consistent REST conventions with versioning to minimize breaking changes.

NFR-603 — Documentation

Architecture, APIs, prompts, and deployment procedures shall be documented and maintained throughout development.

Observability

NFR-701 — Logging

Application services shall generate structured logs for:

- Requests
- Errors
- AI processing
- Background jobs

NFR-702 — Monitoring

The platform shall monitor:

- Response time
- Error rates
- CPU and memory usage
- AI latency
- Queue length

NFR-703 — Alerting

Critical production issues shall trigger alerts for:

- High error rates
- AI service failures
- Processing queue delays
- Infrastructure outages

AI-Specific Requirements

NFR-801 — Explainability

Every AI-generated decision shall include:

- Confidence score
- Supporting evidence from the transcript

NFR-802 — Deterministic Output Format

AI responses shall conform to a structured schema to ensure reliable downstream processing.

NFR-803 — Human Oversight

No AI-generated decision or action item shall become an official organizational record without user approval.

NFR-804 — AI Quality Thresholds

The MVP shall target:

Metric	Target
Decision Precision	≥ 90%
Decision Recall	≥ 85%
Action Item Precision	≥ 90%
Hallucination Rate	≤ 5%

NFR-805 — Model Independence

The architecture shall support replacing the underlying LLM with minimal application changes, reducing vendor lock-in and enabling future optimization.

Operational Targets

Category	Target
Application Response Time	< 500 ms
AI Processing	≤ 30 s (typical transcript)
Search Response	< 1 s
Platform Availability	≥ 99.9%
Concurrent Users	≥ 500
Decision Precision	≥ 90%
Hallucination Rate	≤ 5%
First-Time User Onboarding	≤ 15 min

Validation Strategy

Each non-functional requirement will be validated through:

Requirement	Validation Method
Performance	Load and performance testing

Requirement	Validation Method
Reliability	Failure recovery testing
Security	Security review and penetration testing
Accessibility	Accessibility audits and manual testing
AI Quality	Evaluation datasets and human review
Observability	Monitoring and alert verification

Risks & Mitigations

Risk	Mitigation
Slow AI responses	Background processing, prompt optimization, asynchronous workflows
Rising infrastructure costs	Model routing, caching, usage monitoring
AI service outages	Retry mechanisms and fallback providers
Security breaches	Encryption, least-privilege access, audit logging
Scaling bottlenecks	Cloud-native architecture with horizontally scalable services

Summary

The non-functional requirements establish the quality standards that DecisionFlow must meet to deliver a dependable AI-powered product. By defining measurable expectations for performance, reliability, scalability, security, privacy, accessibility, maintainability, observability, and AI operations, these requirements ensure that the MVP is not only functional but also production-ready and capable of supporting future growth.

Architecture Decisions and Constraints

Decision / constraint	Implication
Asynchronous AI processing	Uploads create jobs; UI polls or subscribes to status rather than blocking
AI service separated from application logic	Models and prompts can evolve without rewriting core domain services
Approved outputs are canonical knowledge	Draft AI objects and verified records require separate states and queries
Tenant-scoped authorization	Every query and storage operation must enforce workspace boundaries
Relational source of truth	Users, meetings, decisions, actions and audit events live in PostgreSQL
Vector retrieval is an expansion capability	MVP can begin with structured filters and approved-text search; semantic search is added when needed

Security and Privacy Review Checklist

- Threat model covers transcript upload, model provider calls, cross-tenant access, prompt injection and data exfiltration.
- Customer data is encrypted in transit and at rest; secrets are stored in managed secret infrastructure.
- Provider contracts and settings prohibit training on customer transcripts without explicit consent.
- Retention and deletion propagate to transcripts, derived outputs, embeddings, logs and backups according to policy.
- Audit events capture approvals, edits, permission changes and administrative actions.
- Sensitive transcript content is excluded from application logs and error telemetry.
- Rate limits, file-size limits, malware scanning and schema validation protect ingestion endpoints.

Target review

Performance, scale and uptime values are engineering targets. They require load tests, cost modeling and operational evidence before they can be represented as achieved service levels.