System Architecture Diagram

Overview

The AI Sales-Enablement Platform follows a microservices architecture designed for scalability, privacy, and seamless integration with existing sales infrastructure. The system processes sales interactions (calls, emails, CRM data) through an intelligent pipeline that generates actionable insights while maintaining complete data privacy.

System Architecture

```
graph TB
    %% External Data Sources
    subgraph "Data Sources"
        CRM[CRM Systems<br/>Salesforce, HubSpot]
        EMAIL[Email Platforms<br/>Outlook, Gmail]
        VOICE[Voice Platforms<br/>Zoom, Teams, Gong]
        MANUAL[Manual Uploads<br/>Documents, Notes]
    end

    %% Integration Layer
    subgraph "Integration Layer"
        API_GW[API Gateway<br/>Rate Limiting, Auth]
        WEBHOOKS[Webhook Handlers<br/>Real-time Events]
        CONNECTORS[Platform Connectors<br/>OAuth, API Clients]
    end

    %% Data Pipeline
    subgraph "Data Processing Pipeline"
        INGEST[Data Ingestion Service<br/>Queue Management]
        PREPROCESS[Preprocessing Engine<br/>Transcript Cleaning, NLP]
        EMBED[Embedding Service<br/>Vector Generation]
        VECTOR_DB[(Vector Database<br/>Pinecone/Weaviate)]
        METADATA_DB[(Metadata Store<br/>PostgreSQL)]
    end

    %% LLM Layer
    subgraph "LLM Hosting Layer"
        LLM_ROUTER[LLM Router<br/>Load Balancing]
        LLAMA_CLUSTER[Llama 3.1 Cluster<br/>LoRA Fine-tuned Models]
        MODEL_STORE[(Model Registry<br/>LoRA Adapters)]
        INFERENCE[Inference Engine<br/>vLLM/TensorRT-LLM]
    end

    %% Application Layer
    subgraph "Application Services"
        INSIGHT_ENGINE[Insight Generation<br/>RAG + Fine-tuning]
        RECOMMENDATION[Recommendation Engine<br/>Next Best Actions]
        ANALYTICS[Analytics Service<br/>Performance Metrics]
        FEEDBACK[Feedback Loop<br/>Learning Pipeline]
    end

    %% Monitoring & Ops
    subgraph "Monitoring & Operations"
        METRICS[Metrics Collection<br/>Prometheus]
        LOGS[Centralized Logging<br/>ELK Stack]
        ALERTS[Alerting System<br/>Model Drift Detection]
        DASHBOARD[Ops Dashboard<br/>Grafana]
    end

    %% User Interface
    subgraph "User Interface"
        WEB_APP[Web Application<br/>React Dashboard]
        MOBILE[Mobile App<br/>iOS/Android]
        API[REST/GraphQL APIs<br/>Third-party Integration]
    end

    %% Data Flow Connections
    CRM --> CONNECTORS
    EMAIL --> CONNECTORS
    VOICE --> CONNECTORS
    MANUAL --> API_GW
    CONNECTORS --> API_GW
    WEBHOOKS --> API_GW
    API_GW --> INGEST

    INGEST --> PREPROCESS
    PREPROCESS --> EMBED
    EMBED --> VECTOR_DB
    PREPROCESS --> METADATA_DB

    VECTOR_DB --> INSIGHT_ENGINE
    METADATA_DB --> INSIGHT_ENGINE
    INSIGHT_ENGINE --> LLM_ROUTER
    LLM_ROUTER --> LLAMA_CLUSTER
    LLAMA_CLUSTER --> MODEL_STORE
    INFERENCE --> LLAMA_CLUSTER

    INSIGHT_ENGINE --> RECOMMENDATION
    RECOMMENDATION --> ANALYTICS
    ANALYTICS --> FEEDBACK
    FEEDBACK --> MODEL_STORE

    INSIGHT_ENGINE --> WEB_APP
    RECOMMENDATION --> WEB_APP
    ANALYTICS --> WEB_APP

    WEB_APP --> API
    MOBILE --> API
    API --> INSIGHT_ENGINE

    %% Monitoring Connections
    LLAMA_CLUSTER -.-> METRICS
    INSIGHT_ENGINE -.-> METRICS
    VECTOR_DB -.-> METRICS
    METRICS --> DASHBOARD
    LOGS --> DASHBOARD
    ALERTS --> DASHBOARD

    %% Styling
    classDef dataSource fill:#e1f5fe
    classDef integration fill:#f3e5f5
    classDef processing fill:#e8f5e8
    classDef llm fill:#fff3e0
    classDef application fill:#fce4ec
    classDef monitoring fill:#f1f8e9
    classDef ui fill:#e3f2fd

    class CRM,EMAIL,VOICE,MANUAL dataSource
    class API_GW,WEBHOOKS,CONNECTORS integration
    class INGEST,PREPROCESS,EMBED,VECTOR_DB,METADATA_DB processing
    class LLM_ROUTER,LLAMA_CLUSTER,MODEL_STORE,INFERENCE llm
    class INSIGHT_ENGINE,RECOMMENDATION,ANALYTICS,FEEDBACK application
    class METRICS,LOGS,ALERTS,DASHBOARD monitoring
    class WEB_APP,MOBILE,API ui
```

Component Descriptions

Data Sources Layer

- **CRM Systems**: Integration with Salesforce, HubSpot, and other CRM platforms
- **Email Platforms**: Connection to Outlook, Gmail for email analysis

- **Voice Platforms**: Integration with Zoom, Teams, Gong for call transcript processing
- **Manual Uploads**: Support for document and note uploads

Integration Layer

- **API Gateway**: Centralized entry point with authentication, rate limiting, and routing
- **Webhook Handlers**: Real-time event processing from external platforms
- **Platform Connectors**: OAuth-based connectors for secure third-party integrations

Data Processing Pipeline

- **Data Ingestion**: Queue-based system for handling high-volume data streams
- **Preprocessing Engine**: Transcript cleaning, NLP preprocessing, and data normalization
- **Embedding Service**: Vector generation using state-of-the-art embedding models
- **Vector Database**: High-performance vector storage for similarity search
- **Metadata Store**: Relational database for structured data and relationships

LLM Hosting Layer

- **LLM Router**: Intelligent load balancing across model instances
- **Llama 3.1 Cluster**: Horizontally scaled deployment of fine-tuned models
- **Model Registry**: Version control and storage for LoRA adapters
- **Inference Engine**: Optimized inference using vLLM or TensorRT-LLM

Application Services

- **Insight Generation**: RAG-enhanced LLM for generating sales insights
- **Recommendation Engine**: AI-powered next best action suggestions
- **Analytics Service**: Performance tracking and business intelligence
- **Feedback Loop**: Continuous learning from user interactions and outcomes

Monitoring & Operations

- **Metrics Collection**: Comprehensive system and model performance monitoring
- **Centralized Logging**: Structured logging for debugging and audit trails
- **Alerting System**: Proactive monitoring for model drift and system issues
- **Operations Dashboard**: Real-time visibility into system health and performance

Data Flow Architecture

1. **Ingestion**: External data sources push/pull data through the integration layer
2. **Processing**: Raw data is cleaned, preprocessed, and converted to embeddings
3. **Storage**: Vectors and metadata are stored in optimized databases
4. **Inference**: User queries trigger RAG-enhanced LLM inference
5. **Response**: Generated insights are delivered through various interfaces
6. **Learning**: User feedback and outcomes feed back into the training pipeline

Security & Privacy

- **Data Isolation**: Complete on-premises deployment with no external data sharing
- **Encryption**: End-to-end encryption for data in transit and at rest
- **Access Control**: Role-based access control with audit logging

- **Compliance**: GDPR, SOC2, and industry-specific compliance support