**AI Recommendation System Sales Offering**

**System Requirements**

**Hardware (On-Premises or Equivalent Cloud Instances)**

**GPUs**

* **Premium Tier**: 4× NVIDIA H100 (80 GB VRAM each)
* **Standard Tier**: 4× NVIDIA A100 (80 GB VRAM each)

**CPU**

* AMD EPYC or Intel Xeon (minimum 16 cores)

**Memory**

* 256 GB+ DDR5 (recommended)

**Storage**

* SSD storage recommended (2TB for fast data access)

**Networking**

* Minimum of 10 Gbps bandwidth recommended

**Hardware Cost Breakdown**

| **Component** | **Recommended Specifications** | **Estimated Cost** |
| --- | --- | --- |
| GPUs | Premium: 4× NVIDIA H100 (80 GB each)Standard: 4× NVIDIA A100 (80 GB each) | Premium: $112k–$146k total  Standard: $40k–$60k total |
| CPU | AMD EPYC or Intel Xeon (≥16 cores) | $5k–$10k |
| RAM | 256 GB+ DDR5 (recommended) | $4k–$8k |
| Storage | 2 TB NVMe SSD | $1k–$3k |
| Networking | 10 Gbps switches/adapters | $2k–$5k |

**Estimated Total Hardware Costs**

* **Premium Tier**: $124k–$172k
* **Standard Tier**: $52k–$86k

**Software, Installation, and Integration Costs**

* **Premium Tier**: $37k–$52k
* **Standard Tier**: $16k–$26k

**Total Solution Cost**

* **Premium Tier**: $161k–$224k
* **Standard Tier**: $68k–$112k

*Note: For cloud deployments, specifications translate to equivalent GPU VM instances (e.g., AWS p4d, Azure ND A100).*

**Software Included (Fully Open-Source Stack)**

**Large Language Model (LLM)**

* **Mistral 7B**: Text generation, summarization, real-time recommendations.

**Embedding & Retrieval**

* **Ollama**: Local embedding generation.
* **ChromaDB**: Vector database for fast similarity searches.

**APIs & Backend**

* **Flask**: Python-based microservice for query processing.
* **Redis**: Optional caching layer.

**Containerization & Orchestration**

* **Docker & Kubernetes**: Application containerization and management.

**Data Ingestion Capabilities**

* **PDF Ingestion**: Automatic extraction and processing of PDF documents
* **URL Ingestion**: Web scraping and content extraction
* **CSV Ingestion**: Structured data processing and integration

**Built-in Applications**

* **AI Recommendation Site**: Web interface for accessing recommendation engine
* **Reporting System**: Comprehensive analytics and reporting tools
* **Chat Interface**: Conversational AI interface for natural interactions
* **Database Management**: Automated maintenance and optimization
* **Metrics Dashboard**: Real-time performance monitoring and analytics

*No recurring vendor license fees; you fully own your infrastructure and software stack.*

**Deployment Options**

**On-Premises Deployment**

* Hardware owned by you.
* Our services include assistance with Kubernetes cluster setup, GPU configuration, and on-site integration.

**Cloud Deployment (AWS/Azure/GCP)**

* GPU VM instances (e.g., AWS p4d, Azure ND A100 v4).
* Terraform and Helm scripts provided for simple, automated provisioning.
* Capability for autoscaling based on workload demands.

Both deployment options deliver the same features and software stack.

*Note: On-premises provides complete data control, while cloud deployments offer easier scalability and management.*

**Services & Support Included**

**Setup & Integration**

* Installation, configuration, system integration.

**Monitoring & Maintenance**

* Regular software updates, troubleshooting support.

**Training & Workshops**

* Workshops for your team on system operation, maintenance, and customization.

**Optional Enhancements**

* Redis integration (optional caching layer).
* Additional custom integrations as per client requirements.