Statement of Work: Rice Market AI System

An Enterprise Resource Planning (ERP) Solution with Natural Language SQL, RAG, and Time-Series Forecasting

Title and Authors:

Title: Rice Market AI System - Enterprise Resource Planning with Natural Language SQL, RAG, and Time-Series Forecasting

Authors:

- Thanh Phong Le (phl690@g.harvard.edu)
- Davar Jamali (davarjamali@gmail.com)
- Pranab Nepal (pranabnepal@gmail.com)

Background and Motivation

Small and Medium-sized Enterprises (SMEs) face high costs for AI/ML ERP systems. We aim to deliver AI microservices that combine transactional data, documents, and time-series pricing to address MLOps challenges.

Problem Statement

Develop a production-ready microservices ERP system enabling SMEs to interact with business data through natural language queries, extract insights from unstructured documents, and forecast prices using AI services.

Data Sources

Dataset 1: Rice Market Pricing Data

- **Source:** Public APIs and repositories (USDA, FAO, commodity exchanges)
- **Description:** Time-series pricing data spanning 5 years across multiple rice varieties
- **Key Attributes:** Date, variety, grade, price per metric ton, moisture content, exchange rates
- **Relevance:** Essential for training LSTM/Prophet ensemble models for price forecasting
- **Data Quality Concerns:** Missing weather correlations will be interpolated

Dataset 2: Rice Mill Operations Database

- **Source:** Production data from partner rice mill (11,000+ rows, Excel format)
- **Description**: Daily operations including production, warehouse, contracts, quality metrics
- **Key Attributes**: Transaction IDs, product codes, quantities, dates, customer information, quality grades
- **Relevance**: Powers NL+SQL service for natural language queries and RAG for document insights
- **Data Quality Concerns:** PII will be scrubbed using Cloud DLP API; Excel formats will be normalized for PostgreSQL ingestion

Scope and Objectives

Primary Objective: Build a scalable, secure ERP system enabling natural language interaction with business data, intelligent document processing, and predictive analytics for rice market operations.

Specific Goals:

1. Natural Language to SQL service translating conversational queries into secure database operations with sub-second response times
2. RAG system for summarizing and extracting insights from business documents using Vertex AI Vector Search

3. Time-series forecasting engine providing 6-month price predictions with explainable confidence intervals using ensemble models

Minimum Components Addressed:

- **Large/Heterogeneous Data**: 5+ years of time-series data, 11,000+ records, unstructured documents
- **Scalability**: GKE with HPA supporting 100+ concurrent users, Cloud SQL read replicas for queries
- **Complex Models**: Fine-tuned Gemini 1.5 Pro for NL understanding, LSTM networks for forecasting, cross-encoder models for document ranking
- **Computationally Expensive Inference**: Vertex AI Vector Search for millions of embeddings, GPU-accelerated transformers, SHAP explainability calculations

Learning Emphasis:

The project leverages course concepts including containerization (Cloud Run), orchestration (GKE), ML pipelines (Vertex AI), and monitoring (Cloud Operations). We'll implement established research: Seq2SQL for database interfaces, RAG foundations from Facebook AI, and Prophet/LSTM architectures for time-series forecasting.

Application Mock Design:

Three primary interfaces deployed on Cloud Run:

1. **Conversational Query Interface**: Natural language input box with real-time SQL translation display and results visualization
2. **Document Analysis Dashboard**: Drag-and-drop upload area, processing status indicators, and extracted insights panel with citations
3. **Forecasting Visualization**: Interactive charts showing 6-month predictions, confidence intervals, and drill-down capabilities for price drivers.

Research and Development:

- Seq2SQL papers (Zhong et al., 2017) for neural database interfaces
- RAG foundations (Lewis et al., 2020) for retrieval-augmented generation
- Prophet (Taylor & Letham, 2018) and LSTM architectures for time-series forecasting
- Vertex AI documentation for managed ML services optimization

Fun Factor

Combining AI with real-world impact for SMEs makes this project engaging. The modular architecture ensures each team member owns specific components while contributing to solving actual business challenges.

Limitations and Risks

- **Technical Risks**: LLM response latency mitigated through Memorystore caching and model optimization
- **Data Quality**: Automated validation pipelines in Cloud Dataflow with quality scoring
- **Model Accuracy**: Fallback mechanisms to simpler models if accuracy drops below thresholds
- **Infrastructure Costs**: GCP free tier optimization and auto-scaling policies to control expenses

Milestones

1. **Project Proposal & Team Formation (09/25)**: Finalize SOW, validate data access, preliminary architecture
2. **MLOps Infrastructure (10/16)**: GCP setup, containerize services, Vertex AI pipeline implementation
3. **Midterm Presentation (10/28)**: Demonstrate NL+SQL (>80% accuracy), functional RAG system, 6-month forecasting
4. **Full-Stack Development (11/25)**: React frontend, REST APIs, CI/CD pipeline, 50%+ test coverage
5. **Deployment & Scaling (12/11)**: GKE deployment, 70% test coverage, demo video, Medium blog post

Rice Market AI System

Enterprise Resource Planning with AI Capabilities