Contact

dbdeveloperexpert@gmail.com

www.linkedin.com/in/milorad-spasic
www.datameshsync.info/portfolio (Portfolio)

Top Skills

Artificial Intelligence (AI)
Retrieval-Augmented Generation (RAG)
Large Language Models (LLM)

Languages

English (Professional Working)
Russian (Professional Working)
Macedonian (Full Professional)

Certifications

HP ASE – Vertica Big Data Solutions Administrator V1
Data Engineering
Microsoft Certified Solutions Expert Data Management and Analytics
Microsoft Certified Solution Expert Business Inteligence
Designing Business Intelligence Solutions with Microsoft SQL Server

Honors-Awards

member of MENSA

Milorad Spasić

Solution Architect ‑ Senior Data Engineer / Cloud & Integration Specialist | PL/SQL developer

Belgrade, Serbia

Summary

Experienced AI Engineer specializing in Large Language Model (LLM) development and deployment with strong expertise in transformer architectures, fine-tuning techniques (LoRA, QLoRA) and Retrieval-Augmented Generation (RAG) systems..

Solution Architect with over 10 years of experience in data engineering and architecture, delivering end-to-end solutions that support real-time analytics, big data processing, and seamless cloud/on-premise integration.

Senior Data Engineer now responsible for defining high-level data strategies, designing reference architectures and guiding engineering teams in implementing secure and maintainable systems.

Core Expertise:

• LLM Engineering & Prompt Design: Practical experience with OpenAI, LLaMA3, Gemma, and local model deployment via Ollama. Skilled in designing efficient prompts and dynamic AI workflows for document analysis, search, and decision-making tasks.

• RAG Systems & Vector Search: Built custom Retrieval-Augmented Generation pipelines using Qdrant and various embedding models. Developed FastAPI-based services that perform real-time document querying, intelligent tagging (via KeyBERT), and semantic search.

• Model Fine-Tuning & Optimization: Worked with LoRA, QLoRA and GGUF formats to fine-tune and serve models locally. Created scripts to train and evaluate models in constrained environments including exporting for use with Ollama.

• AWS PCS p3.2xlarge (HPC) and work from shell prompt with SLURM (sbatch,srun,squeue,scancel), Aptainer for build from .def .sif images for fine-tunning LoRA and QLoRA.

• AI Agents & Multi-Agent Systems: Designed LangChain-based multi-agent pipelines for task decomposition, coordination, and reasoning. Focused on agent orchestration for knowledge-intensive tasks.

• NLP & Transformers: Experience with tokenization, embeddings, and transformer architecture (HuggingFace, PyTorch).

• Deployment & APIs: Built RESTful APIs and Streamlit apps to integrate LLMs into end-user applications.

• Data Engineering & Architecture: Proficient in Python, SQL, and ETL/ELT frameworks like Debezium, Kafka, Greenplum to enable real-time data streaming.

Education:

- University of Belgrade
- Bachelor's degree, Computer Programming

Experience

DataMeshSync

AI Engineer / LLM Engineer & Data & AI Integration Services
January 2023 - Present (2 years 5 months)

AI Engineer / LLM Engineer & Solution Architect – Data & AI Integration

Service

AI Engineer / LLM Engineer specializing in applied Large Language Model (LLM) systems, with hands-on experience in building and optimizing intelligent,production-ready AI solutions. My work combines strong foundations in NLP,vector search, and prompt engineering with modern techniques such as fine-tuning, Retrieval-Augmented Generation (RAG), and multi-agent orchestration.

Core Expertise:

- LLM Engineering & Prompt Design: Practical experience with OpenAI, LlaMA3, Gemma, and local model deployment via Ollama.
- RAG Systems & Vector Search: Built custom Retrieval-Augmented Generation pipelines using Qdrant and various embedding models. Developed FastAPI-based services that perform real-time document querying, intelligent tagging (via KeyBERT), and semantic search.
- Model Fine-Tuning & Optimization: Worked with LoRA, QLoRA, and GGUF formats to fine-tune and serve models locally. Created scripts to train and evaluate models in constrained environments, including exporting for use with Ollama.
- AI Agents & Multi-Agent Systems: Designed LangChain-based multi-agent pipelines for task decomposition, coordination, and reasoning. Focused on agent orchestration for knowledge-intensive tasks.
- NLP & Transformers: Experience with tokenization, embeddings, and transformer architecture (HuggingFace, PyTorch). Able to train and evaluate models for intelligent document processing and NLP automation tasks.
- Deployment & APIs: Built RESTful APIs and Streamlit apps to integrate LLMs into end-user applications. Familiar with scalable deployment using FastAPI and real-time interaction with AI services.

Tooling & Frameworks:

- Python, PyTorch, HuggingFace, LangChain
- Qdrant, Ollama, LLaMA3, Gemma, OpenAI, GGUF, LoRA/QLoRA
- FastAPI, REST API, Streamlit, KeyBERT

Smartivo Technologies

Data engineer | DevOps engineer | Python developer
September 2018 - December 2024 (6 years 4 months)

DevOps and Data engineer with deep expertise in designing and maintaining high-performance, real-time data pipelines. Work has focused on GPS tracking systems and telemetry data collection,processing hundreds of millions of records daily for real-time analytics. With a strong foundation in cloud infrastructure, containerization, and data streaming technologies, specialized in optimizing data workflows for high-velocity, high-volume environments.

Key Responsibilities and Expertise

- Database Change Data Capture (CDC) with Debezium
- Implemented CDC pipelines using Debezium to capture real-time.
- Configured Debezium connectors to ensure minimal latency and high reliability in data streaming.
- Data Streaming with Kafka
- Data Processing and Loading
- Built Python-based consumers to process Kafka streams and load processed data into the Greenplum database.
- Microservices Architecture
- Logging and Monitoring.
- Data Analytics Pipeline Projects

Real-Time Analytics Platform

- Objective: Built a platform to process and analyze real-time data for business intelligence.
- Technologies Used: **Debezium, Kafka, Python, Greenplum, Docker, Kubernetes, Graylog, Prometheus, Grafana.**
- Highlights:
  - Achieved sub-second latency for capturing and processing database changes.
  - Designed scalable microservices to handle millions of daily events.

Microservices-Based Data Pipeline

- Objective: Developed a **microservices architecture** for streaming and processing large datasets from databases.
- Technologies Used: **Debezium, Kafka, Python, Docker, Kubernetes**.
- Highlights:
  - Automated deployment and scaling using Kubernetes.
  - Reduced operational overhead by centralizing logs in Graylog containers.
  - Enhanced fault tolerance with Kafka partitioning and replication.

Infobip

Database developer, backend developer, Senior Database Administrator, DWH architect
March 2014 - April 2018 (4 years 2 months)

Responsibilities:

- Managed and optimized high-volume Microsoft SQL Server and PostgreSQL databases across multiple environments.
- Developed and optimized complex stored procedures, triggers, and functions in T-SQL, PL/PGSQL to improve application performance.
- Designed and implemented high-availability solutions, including Always On Availability Groups, replication, and failover clustering.
- Conducted query performance tuning using execution plans, indexing strategies, and resource management techniques.
- Led database security initiatives, ensuring compliance with GDPR, HIPAA, and industry best practices.
- Developed automated backup and recovery strategies, reducing downtime and enhancing business continuity.
- Collaborated with development teams to optimize application queries and database schema for scalability and efficiency.

Database Architect

Responsibilities:

- Designed scalable, high-performance database architectures for enterprise applications.
- Migrated legacy SQL Server databases to PostgreSQL, ensuring minimal downtime and data integrity.
- Developed ETL processes using PostgreSQL FDW for seamless data integration.
- Implemented advanced security measures, including data encryption, row- level security, and access controls.
- Automated database maintenance tasks using Python and Bash scripting.

Combis

Senior Oracle Developer
May 2013 - December 2013 (8 months) PL/SQL Developer & DWH architect

- Complex stored procedures, functions, triggers, and views
- Advanced indexing strategies (Clustered, Non-Clustered, Full-Text, JSON/ GIN Indexes)
- Query tuning and optimization (Execution Plans, Query Store, Hints, Statistics)
- Partitioning strategies for large datasets