Pavan Suresh

Software Engineer

+1 (314) 655-8773 | pavankumarsuresh35@gmail.com | San Francisco, CA|linkedin.com/in/pavankumarsuresh

SUMMARY

Software Engineer with 5 years of experience building backend and AI-powered platforms using Python, FastAPI, and cloud technologies. Specialized in LLM-based systems, RAG pipelines, and real-time data processing using Kafka, Spark, AWS, and GCP. Strong experience in designing scalable APIs, microservices, and distributed systems with focus on performance, reliability, and cost optimization. Hands-on experience deploying production systems using Docker, Kubernetes, and CI/CD pipelines.

WORK EXPERIENCE

Shopify **San Francisco, CA**

Software Engineer – Generative AI & ML Infrastructure **Oct 2024 – Present**

- Improved AI response accuracy by 32% by implementing Python-based retrieval-augmented generation pipelines with contextual merchant data, optimizing prompt engineering and semantic re-ranking for more relevant and actionable outputs.
- Reduced inference latency by 41% by designing asynchronous FastAPI services with request batching, Redis caching, optimized token handling, API Gateway integration, and rate limiting.
- Decreased infrastructure cost by 28% by implementing dynamic batching and response caching strategies, improving LLM utilization efficiency while supporting AI-driven workflows used by 3,000+ active merchants daily.
- Built LLM-powered APIs for merchant workflows using FastAPI and LangChain, and OpenAI APIs, designing REST and GraphQL APIs with JWT authentication, integrated with React.js frontend dashboards using API calls, enabling structured workflows.
- Developed retrieval pipelines using FAISS, embeddings, and hybrid search techniques, combining semantic similarity with keyword ranking to improve contextual understanding and relevance of AI-generated responses across merchant queries.
- Designed agent orchestration systems with tool-calling capabilities, enabling LLMs to trigger backend APIs for tasks like product updates, campaign creation, and analytics insights through structured multi-step reasoning workflows.
- Architected scalable microservices-based systems using Python, following event-driven microservices architecture and service-to-service communication, simplifying AI service deployment and improving deployment flexibility across distributed systems.
- Implemented event-driven pipelines using Kafka and asynchronous workers on GCP, leveraging Redis pub/sub and distributed messaging, enabling real-time processing with reliable service-to-service communication across systems.
- Deployed containerized services using Docker and Kubernetes on GCP cloud infrastructure, ensuring high availability, auto-scaling, and efficient resource utilization, supporting blue-green deployment strategies for production environments.
- Built CI/CD pipelines using GitHub Actions and monitoring with Prometheus and Grafana on GCP, integrating automated testing, logging, and alerting workflows, improving deployment reliability and proactive system monitoring.

Accenture **India**

Software Engineer **Sep 2020 – Dec 2023**

- Migrated legacy SAS-based data workflows to Python and PySpark pipelines, reducing overall data processing latency by 38% and improving pipeline scalability across high-volume enterprise datasets using distributed data processing architecture.
- Optimized distributed Spark jobs through partition tuning and caching strategies, reducing execution time by 42% and enabling faster analytics delivery for data teams working on large-scale reporting systems.
- Improved data quality and validation processes by implementing automated checks and schema enforcement, increasing data accuracy by 31% and reducing downstream reporting inconsistencies across multiple business-critical datasets.
- Built scalable ETL pipelines using Python, PySpark, and SQL, processing structured and unstructured datasets, integrated with Airflow scheduling and batch processing workflows, enabling efficient data transformation and loading.
- Designed and implemented workflow orchestration using Apache Airflow, creating DAGs for scheduling, monitoring, and managing end-to-end data pipelines with logging, alerting, and retry mechanisms across distributed environments.
- Developed complex SQL queries and data transformation logic for analytics use cases, ensuring optimized data retrieval, aggregation performance, and seamless integration with downstream business intelligence and reporting tools.
- Deployed data pipelines on AWS using S3 for storage, Glue for ETL processing, and EMR for distributed computing, leveraging Lambda and SQS for event-driven processing and asynchronous workflows.
- Implemented event-driven data processing workflows using Kafka and Spark Streaming on AWS, enabling real-time ingestion with distributed messaging and fault-tolerant service communication across microservices-based data systems.
- Containerized data processing services using Docker and managed deployments with Kubernetes, ensuring consistent environments, scalability, and high availability, supporting auto-scaling and blue-green deployment strategies in production.
- Built FastAPI-based REST APIs on AWS behind API Gateway with JWT authentication and rate limiting for secure data access and frontend integration.

TECHNICAL SKILLS

Programming & Core Engineering: Python (Primary), SQL, Bash/Shell Scripting, Data Structures & Algorithms, Object-Oriented Programming (OOP), Design Patterns, Multithreading & Concurrency, Clean Code Practices

Backend Development & APIs: FastAPI, Flask, Django, REST APIs, GraphQL, API Design, Open API/Swagger, Authentication (JWT, OAuth2), Async Programming (async/await), High-Performance API Development, API Gateway Pattern, Rate Limiting

AI/ML & Generative AI: LLMs, Retrieval-Augmented Generation (RAG), Prompt Engineering, LangChain, OpenAI APIs, Embeddings, Semantic Search, Re-ranking, AI Agents & Tool Calling, Model Optimization (Latency, Accuracy, Cost)

Data Engineering & Processing: PySpark, Apache Spark, ETL Pipelines, Data Transformation, Data Validation, SQL Optimization, Data Warehousing, Schema Design, Batch & Stream Processing

Streaming & Event-Driven Systems: Apache Kafka, Spark Streaming, Event-Driven Architecture, Asynchronous Processing, Distributed Messaging Systems, Real-Time Data Pipelines

Databases & Storage: PostgreSQL, Redis, S3, Data Modeling, Query Optimization, Caching Strategies, Vector Databases (FAISS), Hybrid Search (Keyword + Semantic), MySQL, MongoDB, SQL Alchemy, Django ORM, Indexing

Frontend: React.js, TypeScript, JavaScript (ES6+), Next.js, Redux, Tailwind CSS, API Integration (Axios/Fetch)

Cloud, DevOps & Infrastructure: AWS (S3, EMR, Glue, Lambda, API Gateway, SQS, SNS), GCP (Compute Engine, GKE, Cloud Storage), Docker, Kubernetes, CI/CD (GitHub Actions, Jenkins), Microservices Architecture, Distributed Systems, Auto-scaling, High Availability Systems, Monitoring (Prometheus, Grafana, CloudWatch, ELK Stack), Blue-Green Deployment

Soft Skills: Problem Solving, Communication, Collaboration, Ownership, Stakeholder Management, Analytical Thinking, Adaptability, Time Management, Attention to Detail

EDUCATION

Master of Science in Information Systems
Saint Louis University

CERTIFICATIONS

AWS Certified Developer - Associate

PROJECTS

AWS Document Processing Pipeline

- Built an asynchronous document processing system using Python, FastAPI, S3, SQS, and Docker, enabling scalable, non-blocking APIs with reliable background processing.
- Designed secure and fault-tolerant cloud architecture with IAM roles, VPC isolation, retry logic, and DLQ handling, maintaining high availability, data protection, and zero message loss.

ForgeEd – AI-Enhanced Learning Management System

- Developed an AI-powered tutoring assistant using Python, Flask, and GPT APIs, delivering personalized learning support and improving academic query resolution efficiency.
- Built intelligent learning roadmap, assessment, and analytics modules with real-time dashboards, enabling personalized recommendations and data-driven student performance tracking.