VIKAS PABBA

210-873-7446 | pabba.vikas54@gmail.com | https://www.linkedin.com/in/pabbavikas | https://vikas54-7.github.io

SUMMARY

Senior Software Engineer with 4+ years building scalable big data systems processing 50M+ records daily. Expert in Spark, Airflow, and Kafka with proven track record optimizing query performance by 60% and achieving 99.9% uptime. Strong backend development with Spring Boot APIs, SOLID principles, and multi-cloud expertise (AWS, GCP, Azure).

TECHNICAL SKILLS

Languages & Devops: Python, Scala, PySpark, SQL, Java, Docker, Kubernetes, Terraform, CI/CD.
Big Data/Streaming: Apache Spark (PySpark, Scala), Apache Airflow, Apache Kafka, Apache Druid, Hadoop.
Cloud Platforms: AWS (EMR, S3, Glue), GCP (Dataproc, BigQuery), Azure (Databricks, Event Hubs).
Query Engines & Databases: Trino, StarRocks, BigQuery, PostgreSQL, Delta Lake, Redis.
Certifications: Big Data Engineering Bootcamp with GCP and Azure Cloud.

PROFESSIONAL EXPERIENCE

Software Engineer Intern: Websol.ai | Remote | San Antonio, TX Aug 2025 -- Present

- Architected production data pipelines using **Apache Spark (Scala/PySpark) on AWS EMR** processing **50K+ daily records** with **99.9% uptime**.
- Built **RESTful APIs with Spring Boot** serving **10K+ requests/day** with p99 latency <200ms using **SOLID principles** and **Resilience4j**.
- Implemented **real-time streaming with Apache Kafka** and **Apache Airflow** orchestration with **exactly-once semantics**.
- Optimized **Trino, StarRocks, PostgreSQL** queries by 60**%** through indexing, query rewriting, and materialized views.

Graduate Student & Research Assistant | Texas A&M University- Kingsville Aug 2023 – May 2025

- Completed M.S. in Computer Science (GPA: 3.7/4.0) specializing in **Big Data Analytics and Distributed Systems**.
- Conducted research on Spark optimization achieving **70% query performance** improvement through custom partitioning.
- Built production projects with **Airflow DAGs, Kafka streaming, and Kubernetes** auto-scaling on AWS, GCP, Azure.
- Implemented Apache Druid & Trino achieving <100ms OLAP query latency; achieved 85%+ test coverage with ScalaTest.

HCL, India: Data Engineer | HCL Technologies (USAA Client) | Chennai, India July 2021 – July 2023

- Designed **10+ enterprise ETL pipelines** using IBM DataStage and Apache Spark for Fortune 150 banking operations.
- Led migration of **3M+ banking records** with zero data loss and full regulatory compliance using CDC/HVR.
- Optimized batch **jobs by 45%** through parallel processing and Spark tuning on Hadoop clusters (1TB+ daily volumes).
- Built PySpark transformations with **98%+ data quality**, reducing incident **resolution time by 30%**

EDUCATION

Master's in Computer and Information Sciences – Texas A&M University, Kingsville | Aug 2023 – May 2025
Bachelors in Electronics and Communication Engineering – JNTU, India | Aug 2017 – May 2021

ACADEMIC PROJECTS

IoT Smart City Data Lake (Azure + GCP) | PySpark, Airflow, Druid, Trino, Delta Lake

Built IoT platform processing 2M+ events/day using PySpark on Azure Databricks and GCP Dataproc. Implemented medallion architecture with Delta Lake, Airflow orchestration (88 tasks), Apache Druid (<100ms queries), and Trino for federated queries across Azure/GCP/PostgreSQL. Developed ML models achieving 92% accuracy. Deployed with Terraform achieving 35% cost reduction.

E-Commerce Fraud Detection (AWS + Azure) | Spark, Kafka, StarRocks, Spring Boot, Kubernetes

Built fraud detection system processing 100K+ transactions/hour using Scala and Spark Streaming on AWS EMR and Azure Databricks. Implemented Kafka (3 brokers, 12 partitions) and StarRocks reducing query latency to 200ms. Multi-layer fraud scoring achieving 94% precision, 87% recall. Deployed on Kubernetes with HPA auto-scaling (3-10 replicas). 85%+ test coverage.

NYC Taxi Analytics (GCP + AWS) | Spark, Kafka, Airflow, Druid, Trino, Spring Boot

Built platform processing 3M+ daily trips using Spark on GCP Dataproc. Implemented Kafka streaming (3 brokers, 24 partitions), Apache Druid (<100ms OLAP latency), and Trino for federated queries across BigQuery/S3. Orchestrated 50+ Airflow workflows. Deployed Spring Boot APIs on GKE with 99.95% uptime and HPA auto-scaling (3-25 pods). Terraform infrastructure deployment.