VIKAS PABBA

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

SUMMARY

Data Engineer with 3+ years of experience designing and implementing scalable ETL/ELT pipelines in enterprise and regulated environments. Proven expertise in building production-grade data solutions using IBM DataStage, PySpark, Scala, and AWS that process 5TB+ daily data volumes. Specialized in real-time data replication, legacy system migration, and cloud data platform development. Strong track record of optimizing data workflows, reducing processing time by 35%, and ensuring 99.9% data integrity for financial services clients including USAA.

TECHNICAL SKILLS

Languages: Python, Scala, PySpark, SQL, Java | ETL/Big Data: IBM DataStage, Kafka, Spark, Hadoop, CDC, HVR, Airflow.
Cloud & Databases: AWS (S3, EMR, Glue, Lambda), PostgreSQL, MySQL, Oracle, Snowflake.
Tools: Docker, Git, Splunk

PROFESSIONAL EXPERIENCE

Software Engineer Intern: Websol.ai Aug 2025 -- Present

- Build data pipelines with Scala/Python/PySpark on AWS EMR processing **500K+ daily records** with sub-minute latency.
- Implement real-time streaming using Apache Kafka for event-driven architecture supporting business analytics.
- Optimize SQL queries improving PostgreSQL **performance by 40%** through indexing and partitioning strategies.
- Create automated data quality **frameworks ensuring 99.9% data integrity** across all pipelines.

HCL, India: Security Integrated ETL Developer (USAA Client) Jan 2022 – July 2023

- Architected 25+ enterprise ETL pipelines using IBM DataStage 11.x processing banking data for 13M+ members.
- Led **10M+ record migration** from legacy mainframe to modern platforms with full audit compliance and zero data loss.
- Implemented real-time replication using CDC/HVR, reducing data latency from hours to minutes.
- Optimized DataStage jobs **achieving 35% runtime improvement**, reducing batch time from 6 hours to under 4 hours.
- Developed PySpark transformations on **Hadoop processing 5TB+ daily volumes** with distributed computing patterns.

HCL, India: Data Migration Developer Jun 2021 – Dec 2021

- Executed end-to-end migration of **100K+ healthcare records with HIPAA-compliant encryption** and zero data loss
- Designed ETL processes using DataStage and Python for extraction, transformation, validation, and loading.
- Implemented field-level data masking and created comprehensive audit trails for regulatory compliance

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

Real-Time NYC Taxi Analytics Pipeline | PySpark, Kafka, AWS, Scala

Architected end-to-end streaming pipeline processing **2M+ monthly records** with Kafka producers (1000 events/sec), PySpark jobs on AWS EMR calculating demand patterns and fare analytics, Scala-based anomaly detection identifying 500+ suspicious trips, 15+ automated data quality checks, and S3 storage with Parquet compression achieving **40% cost reduction** and **sub-minute latency.**

E-Commerce Data Warehouse | Python, PySpark, PostgreSQL, Airflow

Designed star schema warehouse with 3 fact tables and 5 dimensions processing **100K+ orders**. Built Python ETL with incremental CDC patterns reducing processing by 70%, PySpark transformations for customer lifetime value and cohort analysis, 20+ data quality validation rules, Airflow orchestration achieving **99.9% reliability**, and PostgreSQL optimization with indexing reducing query time **45%** (8s to 4.4s).

Stock Market Data Platform | Scala, Spark Streaming, Kafka, AWS S3

Developed Scala streaming application consuming **500+ stock symbols** every 5 minutes via Alpha Vantage API with rate-limiting and retry logic. Implemented Kafka topics by sector (10K msg/sec), windowed aggregations for moving averages and technical indicators, data enrichment with company fundamentals, partitioned S3 storage (**65% compression**), and anomaly detection with **98% accuracy.**