Satyam Kumar

+91 7042746946 | satyamkumar2107@gmail.com | LinkedIn

PROFESSIONAL SUMMARY

Data Engineer with 3 years of experience designing scalable ETL and streaming data platforms on AWS and Spark ecosystems. Skilled in building production-grade pipelines using Glue, PySpark, Airflow, Kafka, and Redshift with strong focus on reliability, performance optimization, and monitoring. Proven track record of reducing pipeline runtimes by 40 - 65%, improving system stability, and architecting solutions for large-scale enterprise datasets (300M+ records).

Technical Skills

Programming Languages : Python, SQL
Big Data Technologies : PySpark, Spark SQL, Apache Spark, Kafka (Streaming), Hive
Cloud & Data Platforms : AWS S3, Glue, Redshift, EMR, EC2, IAM, CloudWatch, Databricks, ADLS , ADF
Orchestration & DevOps Tools : Apache Airflow, Docker, Kubernetes, Git, CI/CD
Visualization / Reporting: Streamlit
Core Concepts: ETL/ELT, Real-time & Batch Pipelines, Data Modeling, Data Quality, Monitoring & Alerting, Performance Tuning, Production Support

Experience

Concert AI - *Data Engineer* Sep 2025 - Present

- Re-architected ML data processing from notebook-based execution to distributed AWS Glue + PySpark pipelines, reducing runtime by 65% and enabling 3× faster processing.
- Built a Streamlit-based weekly reporting application used by stakeholders and analysts to compare new vs. prior-week datasets, reducing manual analysis effort by 70%+ and enabling faster decision-making.
- Engineered large-scale transformations and joins across 300M+ rows, improving query performance by 40%.
- Designed controller-based workflow orchestration framework in AWS Glue to dynamically trigger dependent jobs with parameter passing, enabling fully automated multi-stage pipelines.
- Refactored legacy scripts into an object-oriented PySpark framework with a reusable base class extended by 30+ jobs, reducing code redundancy and improving maintainability.
- Automated external table creation for S3 datasets, reducing manual setup effort by 80%.
- Consolidated multiple schemas into unified analytics layer, simplifying downstream reporting and improving reliability.
- Built an automated dashboard export solution that generated pre-formatted Excel reports, reducing weekly manual reporting effort by ~95% (from 2+ hours to under 5 minutes).
- Diagnosed architectural flaw in distributed Spark inference pipeline where partitioning caused inconsistent model predictions; redesigned workflow using AWS Step Functions + EC2 single-node execution, restoring prediction accuracy and ensuring deterministic results.

Deloitte - *Data Engineer* Feb 2023 - Sep 2025

Clients - Takeda Pharmaceutical , Vanguard

- Designed and deployed scalable PySpark + AWS Glue ETL pipelines for large-scale transactional/logistics datasets across 10,000+ standardized tables
- Migrated legacy Informatica workflows to AWS-native Spark architecture, improving runtime by 40% and reducing infrastructure costs
- Built an end-to-end ingestion and transformation pipeline to move and prepare data from DynamoDB to S3, helping replace manual paper forms with digital signatures.
- Designed event-driven ingestion patterns enabling near real-time data availability for analytics
- Optimized client data models by reducing schema complexity by ~50%, which improved query performance and reduced storage costs.
- Implemented an automated alerting mechanism using Power Automate to notify team members via Microsoft Teams, upon AWS Glue job failures, reducing reliance on email monitoring and enabling faster incident response.
- Optimized Glue jobs using partition pruning, caching, and Spark tuning to reduce memory usage and execution time.

- Designed and orchestrated 100+ production DAGs using Apache Airflow with retries, SLA alerts, and failure recovery ensuring 99.9% pipeline reliability
- Owned UAT and production releases, ensuring smooth cutovers and minimal downtime
- Independently developed a low-code internal tool using Power Apps + Power Automate to streamline assessments, improving efficiency across business units.

Education

Jain University Bangalore, Karnataka
Master of Computer Applications **(MCA)** March 2023 - May 2025

CMR University Bangalore, Karnataka
Bachelor of Computer Applications **(BCA)** July 2019 - May 2022

Certifications and Awards

- **Databricks** Data Engineer Certificate
- **AWS [** Certified Cloud Practitioner , Certified Solutions Architect – Associate**]**
- Nominated and Awarded **[Spot Award]** for developing "Internal ILA assessment" tool for Deloitte.
- Awarded **[Applause Award]** for outstanding impact as a Data Engineer for driving development resulting in significantly enhanced business data insights, decision-making and cost saving.