Harshith Kumar K

Senior Data Engineer

+919980331889 ◇ harshithkumark118@gmail.com ◇ Bengaluru, Karnataka, India ◇ Open to Relocate ◇ LinkedIn

SUMMARY

Senior Data Engineer with 4+ years of experience in designing scalable data architecture, integrating external data sources, and optimizing data infrastructure for actionable data solutions. Expertise in AWS, Python, SQL, and distributed systems like Apache Spark and Kafka for data manipulation and transformation. Proven ability to enhance backend services, improve query performance, and drive ML-ready data pipelines with automation tools for big data analytics.

WORK EXPERIENCE

Senior Data Engineer Oct '24 — Present
Teksystems Global Solution Bengaluru, India

- Contributed to the migration of **100 TB+** big data analytics from Hadoop to AWS S3, leveraging **AWS Glue (PySpark, SQL)** and **Athena (SQL)** to improve query performance by 50% and enable scalable **data solutions** for actionable **financial data architecture.**
- Worked on ETL pipelines processing structured (Parquet, Avro) and semi-structured (XML) datasets, integrating third party vendor data via **AppFlow** for enhanced data manipulation, transformation, and visualization accessibility.
- Optimized **Athena SQL** queries, reducing processing **time by 50%**, and improved data completeness by integrating missing records for the data science team, enhancing analytical skills and data-driven intelligence for marketing platforms.
- Designed a daily data load pipeline using **Control-M** and **EC2** as a bridge to t**rigger AWS Lambda**, automating batch ingestion into **Athena and Redshift**, reducing operational **costs by 40%.**
- Helped develop real-time **Change Data Capture (CDC)** pipelines, ensuring data consistency across **data lakes and marts**, while leveraging **CloudWatch** for **monitoring and automated alerts.**
- Tested and validated **data pipelines,** improving data integrity and performance, and contributed to comprehensive documentation in **Confluence for streamlined knowledge** sharing.
- Enhanced **CI/CD** pipelines on **GitHub** and managed **JIRA sprints**, increasing deployment speed by **20%.**
- Integrated **AWS Kinesis** for event-driven data processing, optimizing real-time analytics and **ML workloads.**

Data Analytics and Data Engineering Associate Feb '21 — Sep '24
Labcorp Laboratories India Bengaluru, India

- Engineered **scalable ETL pipelines** using **AWS Glue (PySpark, SQL), S3, and Redshift**, reducing **latency from 4 hours to 1 hour** and improving **data reliability** for marketing analytics and digital transformation.
- Developed **predictive models using AWS SageMaker**, achieving **80% forecasting accuracy**, enabling **data-driven decision-making for customer-focused data solutions** and speech analytics.
- Processed and transformed **unstructured data** using **Python, NLP, and AWS Lambda**, achieving **98% data extraction accuracy** and enhancing **event tracking and business intelligence operations**.
- Integrated **AWS Step Functions and Lambda** to orchestrate automated workflows, optimizing data ingestion into **Redshift and Athena**, reducing **operational overhead by 40%**.
- Implemented **real-time Change Data Capture (CDC) pipelines**, ensuring **data consistency across data lakes and marts**, while leveraging **CloudWatch for monitoring and automated alerts**.
- Contributed to **CI/CD automation** using **GitHub Actions** and improved **JIRA sprint workflows**, accelerating **deployment cycles by 20%**.
- Processed and transformed **unstructured data** using **Python, NLP, and AWS Lambda**, achieving **98% data extraction accuracy** and enhancing **event tracking and business intelligence metrics**.

EDUCATION

Master in Computer Applications, R.V. College of Engineering Oct '18 — Jun '21
Bangalore, India

Bachelor in Computer Applications, Yuvaraja's College Jul '15 — Jun '18
Mysore, India

CERTIFICATIONS

AWS Certified Data Engineer Associate Aug '24

AWS Certified Solution Architect Feb '25

PERSONAL PROJECT

Real-Time CDC Pipeline with AWS Iceberg & Athena

- Built a real-time CDC pipeline using AWS Lambda, EventBridge, Glue (PySpark), and Iceberg, enabling efficient data processing and querying in Athena, with insights visualized in QuickSight.

AWS-Based CDC ETL Pipeline for E-Commerce

- Developed a CDC ETL pipeline with Apache Airflow, AWS Glue (PySpark), and Redshift, ingesting PostgreSQL data and optimizing performance with partitioning, indexing, RBAC and PII.

SKILLS

Relevant Skills Communication Skills, Data Solutions, Data Manipulation, Data Architecture, Automation Tools
Cloud Technologies Amazon Web Services, AWS (S3, Glue, Redshift, Lambda, Athena, Sage Maker, Step Function, EC2, EMR)
Data Engineering Big Data, Apache Spark, Data Driven, Data Models, Data Pipelines, ETL/ELT, Apache Airflow
Data Management Data Lakes, Data Warehousing, Big Data Analytics
Programming Languages Python, SQL, PySpark
Analytics Tools Data Visualization, Tableau, Power BI, MySQL, AWS RDS