Mayur Jain

Long Beach, CA | 562-254-7817 | mayurjain333@gmail.com | linkedin.com/in/mayurjain007 | github.com/mayurjainf007

Summary

Data Engineer with 5+ years of experience designing scalable ETL and data pipelines across cloud platforms. Specialized in real-time streaming, distributed processing, and analytics optimization. Proven impact in healthcare and BFSI domains through data-driven solutions. Skilled in Agile leadership and driving strategic decisions through data insights.

Skills

Programming: Python, Java, SQL, PySpark, Git, Bitbucket
Big Data Tools: Apache Spark, Kafka, NiFi, Flink, Hadoop, Hive, Cassandra, Snowflake
DevOps: AWS (Glue, EMR, Redshift), GCP, Docker, Kubernetes, Terraform, Jenkins, CI/CD
Orchestration: Airflow, Dataiku, Azure Data Factory, Databricks
ML & Visualization: MLFlow, TensorFlow, Tableau, Power BI
Other: Data Lakehouse, Data Governance, Unix, Agile, JIRA, Confluence

Professional Experience

Graduate Data Engineering Analyst (Student Assistant) *Mar 2024 - Present*
CSULB 49er Foundation (Non-Profit)

- Automated financial reporting with Python and Excel VBA, reducing manual workload by 40%.
- Analyzed 10K+ transactions using PySpark SQL, delivering insights that improved policy compliance.
- Built interactive Tableau dashboards for executive reporting, enabling KPI-driven budget reviews.

Lead Data Engineer *Sep 2022 - Jul 2023*
ZS Associates

- Built ML pipeline in PySpark & Dataiku to guide HCP targeting by region, doctor, insurer & seasonality.
- Used patient-level data to predict drug-market fit by geography & time, improving ROI by 30%.
- Orchestrated 10+ pipelines on Databricks & Azure DevOps using Git/Bitbucket with automated schema change.
- Ensured 99.9% data integrity using Collibra while analyzing over 100M+ sensitive healthcare records.
- Tuned models & ran A/B tests, boosting targeting by 25% and improving KPI-driven decisions.
- Enforced healthcare data security, maintaining 100% compliance with client confidentiality policies.

Senior Data Engineer *Jul 2019 - Sep 2022*
Tata Consultancy Services

- Built 10+ ETL pipelines in PySpark on Databricks and Redshift to detect fraud across BFSI transactions.
- Modeled financial and consumer data into star schemas for long-term KPI trend reporting via Tableau.
- Deployed market segmentation logic to detect anomalies with a 20% improvement in fraud detection accuracy.
- Maintained 99.5% data integrity and 100% SLA adherence through automated quality checks and lineage tracking.
- Orchestrated real-time alerting using Airflow and Redshift, reducing fraud response time by 35%.

Software Development Engineer *Aug 2017 - Jul 2019*
ClickIndia Infomedia

- Developed APIs with Spark and Databricks for real-time data workflows, eliminating 80% manual effort.
- Tuned SparkSQL over Cassandra to reduce latency by 30% during high-traffic scenarios.
- Optimized infrastructure for cloud performance, cutting latency by 25% across deployments.

Projects

- **Heart Disease Detection:** Streamed patient vitals via Kafka, applied SparkML models for real-time risk scoring.
- **Fraud Detection System:** Built event-driven fraud alerting using Spark & Airflow; reduced false positives.
- **Sentiment Analysis:** Developed NLP pipeline in Airflow using Twitter API to monitor brand perception.

Certifications

- Microsoft Certified: Azure Data Engineer Associate
- AWS Certified: Data Engineer Associate
- Published thought leadership on scalable data platforms and real-time analytics (Medium @mayurjain007).

Education

California State University Long Beach, US *May 2025*
Master of Science in Computer Science
Guru Tegh Bahadur Institute of Technology, IN *May 2019*
Bachelor of Technology in Information Technology