Mayur Jain

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

Summary

Data Engineer with 6+ years of experience designing scalable data pipelines across multi-cloud environments. 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, Go
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

Work Experience

Data Engineering Intern — *Mar 2024 - Present*
CSULB 49er BeachShops
- 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.
- Delivered Tableau-based executive insights monthly, supporting KPI reviews and strategic decisions.

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 file ingestion & 10+ pipelines on Databricks (via Git/Bitbucket) with automated schema validation.
- 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 HIPAA-compliant healthcare data policies; supported code reviews and QA for operational excellence.

Senior Data Engineer — *Jul 2019 - Sep 2022*
Tata Consultancy Services
- Built 10+ ETL pipelines on Databricks and NiFi using PySpark to detect fraud across BFSI transactions.
- Architected hybrid data flows linking on-prem NiFi and cloud Databricks, improving data sync efficiency by 25%.
- Deployed market segmentation logic to detect anomalies with a 20% improvement in fraud detection accuracy.
- Monitored Spark pipeline ingestion with automated checks; ensured 99.5% data integrity & uptime.
- Mentored 5+ junior engineers on governance and led real-time alerting via Airflow, cutting fraud response by 35%.

Software Development Engineer — *Aug 2017 - Jul 2019*
ClickIndia Infomedia
- Developed microservices with Databricks for real-time data workflows, eliminating 80% manual effort.
- Tuned SparkSQL over GCP BigQuery to reduce latency by 30% during high-traffic scenarios.
- Migrated SparkSQL workloads to GCP BigQuery, reducing compute latency by 30% in high-traffic periods.

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 — *Aug 2023 - May 2025*
Master of Science, Computer Science — *California, US*

Guru Tegh Bahadur Institute of Technology — *Aug 2015 - May 2019*
Bachelor of Technology, Information Technology — *Delhi, IN*