Mayur Jain

562-254-7817 | mayurjain333@gmail.com | LinkedIn/mayurjain007 | GitHub/mayurjainf007 | mayurjainf007.github.io/Resume

SUMMARY

Data Engineer with 5+ years of experience building large-scale data ecosystems, production-grade pipelines, and analytics/ML infrastructure across healthcare, consulting, and finance domains. I've engineered high-throughput ingestion frameworks, automated quality controls, deployed predictive models into real workflows, and driven measurable operational and financial impact. I focus on building systems that are reliable, scalable, and actually used, not theoretical showpieces.

SKILLS

- **Programming Languages**: Python, SQL, PySpark, Java, R, Scala
- **ETL Tools**: Apache Kafka, Spark, Airflow, Databricks, Snowflake
- **Methodology**: SDLC, Agile, Waterfall, CRISP-DM
- **IDES**: PyCharm, Jupyter Notebook, IntelliJ IDEA, Visual Studio, NetBeans
- **Big Data Ecosystem**: Hadoop, MapReduce, Hive, Pig, HDFS, Yarn, Data Lake, Data Warehouse
- **Visualization and Reporting Tools**: Power BI, Tableau, SAS, Argos, Looker, Amazon QuickSight, Google Data Studio
- **Data Science & Machine Learning**: Statistical Modeling, Hypothesis Testing, Feature Engineering, Model Training, Model Evaluation, Classification, Regression, Clustering, Time Series, Natural Language Processing (NLP), Deep Learning (Neural Networks, CNNs, RNNs, LSTMs), Recommendation Systems, Model Deployment & Monitoring (MLOps)
- **Cloud Platforms**: AWS (S3, Glue, IAM, Lambda, EC2, Redshift, RDS, CloudWatch, EMR, SageMaker, DynamoDB), Azure, GCP
- **Generative AI**: Large Language Models (LLMs), Prompt Engineering, Text Generation, Image Generation Models, Embedding Models, Vector Search & Retrieval, Retrieval-Augmented Generation (RAG), Fine-tuning & Instruction-tuning LLMs
- **Packages**: NumPy, Pandas, Matplotlib, SciPy, Scikit-learn, Seaborn, TensorFlow
- **Development and Operations Tools**: Git, GitLab, Jenkins (CI/CD), Docker, Kubernetes, Grafana, Splunk, Jira, GitHub
- **Databases**: MySQL, PostgreSQL, MongoDB, Neo4j, RDBMS, NoSQL, Cassandra, HBase, Elasticsearch
- **Operating Systems**: Windows, MacOS, Kali Linux, Ubuntu

EXPERIENCE

Cigna Healthcare **Jul 2025 - Present**
Data Engineer *New Jersey*

- Built a unified healthcare Data Lake on AWS (S3, Glue, Lambda, Redshift, EMR) to consolidate fragmented claims, EHR, and policy data, improved data availability for analytics teams by 60% and eliminated multiple legacy ingestion jobs.
- Developed predictive models in Python/PySpark to detect high-risk members, early chronic-condition indicators, and cost anomalies, integrating these models into operational workflows via SageMaker endpoints with automated model monitoring.
- Implemented RAG pipelines using LLMs and vector search to extract insights from policy documents, clinical guidelines, and claims notes, reducing manual review time for care management teams by 55%.
- Built and automated data quality frameworks for PHI/PII compliance (HIPAA) using PySpark SQL and CloudWatch, catching schema drift, duplicate claims submissions, and data corruption before reaching downstream systems.
- Created dashboards in Power BI/Tableau for Claims Ops and Care Management, visualizing utilization trends, fraud patterns, reimbursement cycles, and provider performance; enabled leadership to identify $8M+ in avoidable spend annually.

Associated Students Inc. **Mar 2024 - May 2025**
Data Engineering Analytics Assistant *California*

- Developed SQL/Python scripts to clean, validate, and transform student data, improving accuracy across campus reporting systems.
- Automated recurring datasets and refreshed analytics workflows, reducing manual reporting time.
- Built dashboards and ad-hoc analyses to support budgeting, staffing, and operational insights for student programs.
- Collaborated across data teams to troubleshoot issues, standardize data definitions, and improve data documentation.

ZS Associates Ltd. **Sep 2022 - Jul 2023**
Associate Consultant (Data Engineer) *India*

- Built high-throughput Python/SQL ingestion pipelines for terabyte-scale oncology RWD/RWE datasets, improving efficiency by 50%.
- Designed oncology-based Snowflake transformation layers and optimized table structures, reducing analytics preparation time by 30%.
- Delivered 25+ analytics-ready datasets for patient insights and commercial reports, resulting in faster and more informed decisions.
- Enhanced reliability of PySpark/SQL pipelines with stronger validation rules and performance tuning, cutting rework by 25%.

Tata Consultancy Services Ltd. **Jul 2019 - Sep 2022**
System Engineer (Data Engineer) *India*

- Built and maintained more than 10 production ETL pipelines for banking datasets using Python, SQL, Hive, and Airflow.
- Automated data validations and anomaly checks in Spark SQL pipelines, reducing manual QA effort across financial workflows.
- Improved reliability of financial data pipelines by optimizing Spark transformations and implementing automated validation checks.
- Optimized Spark-based fraud detection pipelines and reduced decision latency by 35% through efficient feature processing.

EDUCATION

California State University Long Beach **Aug 2023 - May 2025**
Master of Science, Computer Science

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