Sai Varre

Richmond, VA | saivarre98@gmail.com | +1(346)-285-2722 |LinkedIn |Trailhead

Summary

Data Engineer with 4+ years of experience designing, building, and optimizing **batch** and **streaming data pipelines** across **Azure**, **AWS**, and **GCP** environments. Hands-on with **Python**, **SQL**, **Spark**, **Kafka**, **Airflow**, **dbt**, and **Snowflake**, delivering scalable **ETL/ELT workflows** and high-quality analytical data models supporting enterprise reporting. Skilled in implementing **data quality validation**, **schema enforcement**, and **orchestration reliability** using **Great Expectations**, **Kafka Schema Registry**, and parameterized **Airflow** pipelines. Experienced partnering with analytics and ML teams to operationalize **production-ready dataflows**, improve reporting accuracy, and support **predictive** and **anomaly detection** use cases through modern **DataOps practices** and governed data infrastructure.

Work Experience

AT&T — Data Engineer **May 2023 - Present**

Cloud Data Platform | ETL Orchestration | DataOps & Observability | Telecom Analytics

- Engineered and maintained scalable **real-time ETL pipelines** using **Azure Data Factory**, **AWS Glue**, and **Kafka**, processing **10+ TB/month** of telecom data (**Parquet/Delta**) for analytics dashboards.
- Implemented **Spark Streaming** and **Snowflake Streams** frameworks to achieve **sub-minute latency** for network performance analytics and monitoring.
- Contributed to standardized **ingestion + transformation patterns** across **Azure, GCP,** and **Snowflake** environments, improving code reusability and reducing onboarding time for new pipelines.
- Automated **data validation** and **quality checks** using **Great Expectations**, detecting schema drift and null anomalies, and reducing data quality incidents by **35%**.
- Introduced producer–consumer data contracts using **Kafka Schema Registry** and **CI checks** (backward-compatibility, required fields), preventing breaking changes before deploys.
- Developed and automated modular **dbt** transformations for **200+** business tables (with CI/CD), supporting customer **behavior**, fraud detection, and predictive maintenance analytics.
- Deployed cross-cloud pipelines using **GCP Dataflow**, **Cloud Storage**, **BigQuery** and **Cloud Composer**, automating ingestion from **Azure Data Lake** and enabling unified, monitored analytics across environments.
- Configured **Pub/Sub connectors** for near real-time synchronization between **Kafka** and **BigQuery**, improving multi-region data accessibility and query performance by **20%**.
- Added **table-level lineage** and **run-history metadata logging** for batch and streaming pipelines, improving traceability and speeding up root-cause analysis during data incidents.
- Improved **Snowflake** performance and reduced cloud costs by **~20%** through warehouse tuning and optimized query caching.
- Enhanced data observability with **Grafana, Azure Monitor** and **Power BI dashboards** to monitor pipeline throughput, SLA breaches, and error trends across distributed workloads.
- Collaborated with Data Scientists to **operationalize ML pipelines** within **Databricks**, integrating predictive churn and anomaly detection models into production dataflows.
- Created parameterized **Airflow DAGs** to **manage cross-cloud dependencies** among Snowflake, Redshift, and PostgreSQL workloads, improving reliability and automation.
- Enhanced pipeline reliability by introducing f**ault-tolerant execution patterns**, including **retry handling** and **targeted backfills**, which reduced on-call intervention during pipeline failures.
- Partnered with product teams to define **telecom data governance standards**—cataloging, lineage tracking, and access controls—using **Azure Purview** and **AWS Glue Data Catalog**.
- Authored **runbooks, support playbooks,** and **post-incident documentation,** and coordinated **RCA review sessions,** improving onboarding, knowledge transfer, and support handoffs across data teams.

BluePal Solutions Pvt Ltd, India — Data Engineer **Nov 2020 - Aug 2022**

- Built **Python ETL pipelines** to process academic and financial data into **PostgreSQL**, reducing manual reporting work by **40%**.
- Designed **PostgreSQL schemas with indexes and partitions**, cutting query latency from 5s+ to under 2s for high-volume analytics queries.
- Integrated **10+** university systems with secure **REST APIs**, enabling consistent data exchange for enrollment, finance, and student management.
- Automated **ETL testing and validation** with Python, SQL, and JUnit, improving trust in nightly pipelines and reducing defects by **25%**.
- Developed **Kafka producers and consumers** to stream enrollment and attendance events in real time, providing administrators timely operational insights.
- Containerized ETL workloads with **Docker and Jenkins CI/CD**, improving consistency across environments and cutting deployment effort by half.
- Implemented **Python + Great Expectations validation suites** for nightly ETL jobs, reducing reporting errors by **25%** and improving PostgreSQL data reliability.
- Migrated legacy academic data pipelines to **GCP BigQuery (sandbox)** and **Cloud Storage** for benchmarking, introducing cost-efficient analytics at scale.
- Collaborated with BI teams and delivered **5+** forecasting dashboards for enrollment planning (used by **200+** administrators) to improve resource allocation.
- Supported deployment of **containerized ETL services** to **Azure App Service** using **Azure DevOps** pipelines, contributing to the team's early cloud migration efforts.
- Designed **role-based access** and basic **data-anonymization** scripts in Python/SQL to secure sensitive student records.
- Wrote **technical documentation and onboarding guides**, reducing new engineer ramp-up time by **30%** and promoting best practices.
- Monitored production by analyzing **PostgreSQL logs and Kafka streams**, diagnosing bottlenecks and sustaining **99.9%** pipeline uptime.

Technical Skills

Programming Languages: Java, Python, Shell, C, C++, Bash, Scala, SQL, YAML
Data Engineering & Big Data: ETL/ELT, ETL Testing, Apache Kafka, Apache Spark, Hadoop, Airflow, dbt, Delta Lake, Databricks, Great Expectations, DataOps, Data Modeling
Frameworks / Libraries: Flask, TensorFlow, PyTorch, Scikit-learn, spaCy, NLTK, BERT, PySpark
Cloud & DevOps: AWS (S3, Glue, Kinesis, Lambda, CDK, CloudWatch, CodePipeline, CodeBuild), Azure (ADF, Synapse, Azure DevOps, Purview, Monitor), GCP (BigQuery, Dataflow, Pub/Sub, Composer), Docker, Jenkins, Terraform
Databases & Warehousing: PostgreSQL, MySQL, Oracle, MongoDB, Snowflake, Redshift, BigQuery, NoSQL
Testing & Monitoring & Logging: ELK Stack (Elasticsearch, Logstash, Kibana), Datadog, CloudWatch (logs, metrics, alarms)
Data Analysis & Visualization: Tableau, Power BI, Excel (Advanced), Matplotlib, Seaborn
Testing Tools: JUnit, PyTest, JMeter, REST Assured (for pipeline/API testing)
Development & Collaboration Tools: Git, GitHub, Bitbucket, Jira, Confluence, Jupyter, Pandas, NumPy, TKinter

Education

University of Wisconsin-Milwaukee – Master's in Computer Science **May 2024**
Related Coursework: Natural Language Processing, Robot Motion Planning, Machine Learning, Computational Intelligence, Computer Networks, Computational Models Decision Making, Operating Systems, Data Analytics
Jawaharlal Nehru Technological University, Hyderabad – B.Tech in Computer Science Engineering **Nov 2020**