

Sai Shreya Peyyala

Email: saishreya.peyyala@gmail.com Phone: +91-7981849988 GitHub: GitHub LinkedIn: LinkedIn

SUMMARY

Senior Data Engineer with 4+ years of experience designing, building, and maintaining scalable cloud-based data pipelines on GCP and Azure. Hands-on expertise in PySpark, Dataflow, BigQuery, Databricks, and Apache Airflow, with strong experience in CI/CD automation using Jenkins and GitHub. Proven track record of optimizing pipeline performance, improving data quality, supporting governance initiatives, and delivering reliable analytics-ready data in enterprise environments.

SKILLS

Programming & Scripting: Python, SQL, Shell Scripting, Linux

Big Data & Streaming: PySpark, Apache Kafka, Hadoop

Cloud Platforms: GCP (BigQuery, Dataflow, Dataproc, GCS, Composer), Azure (Databricks, Data Factory, Azure Data Lake)

Workflow Orchestration: Apache Airflow

Databases: Relational (Oracle), NoSQL (MongoDB)

DevOps & CI/CD: Git, GitHub, Jenkins, Docker

Data Engineering Concepts: ETL/ELT, Data Quality, Data Governance, Pipeline Monitoring, Cost Optimization

WORK EXPERIENCE

Cognizant, Hyderabad

Mar 2025 – Present

Data Engineer

- Migrated 500+ ETL workflows from Informatica PowerCenter to Google Cloud Dataflow . Worked closely with legacy Informatica mappings, workflows, and scheduling dependencies.
- Designed batch scheduling and dependency management workflows comparable to Autosys using Apache Airflow .
- Implemented reconciliation checks, control totals, and audit-ready data validation for enterprise reporting .
- Supported production issue resolution with root cause analysis and change management .
- Optimized Google Cloud Dataflow pipelines by analyzing long-running tasks, identifying performance bottlenecks, and implementing enhancements that reduced overall pipeline execution time.
- Took ownership of end-to-end ETL modules and guided junior engineers through design, development, and production support activities.
- Designed, built, and maintained end-to-end data pipelines using Apache Airflow for orchestration, scheduling, monitoring, and failure handling.
- **Tools:** Python, SQL, Google Cloud Dataflow, Apache Airflow, Google BigQuery, GitHub, Docker, Shell Scripting

Tata Consultancy Services, Hyderabad

June 2021 – Mar 2025

Data Engineer

- Developed a scalable ETL framework for data migration at PayPal using Python, Google Cloud Storage (GCS), and BigQuery, reducing migration time by 20% and improving scalability by 30%.
- Automated data validation and data quality checks by integrating SQL-based rules into a centralized execution framework.
- Migrated legacy Hive-based ETL workflows to PySpark, reducing runtime by up to 70% and cutting execution costs by over 50%.
- Optimized Spark queries and transformations, achieving performance improvements of up to 70%.
- Designed and deployed CI/CD pipelines using Jenkins, and managed Docker image builds and container registries.
- Built and optimized Azure Data Factory (ADF) ETL pipelines leveraging Medallion (Bronze–Silver–Gold) architecture, ingesting data from 10+ REST APIs into 16+ curated tables and improving processing efficiency by 55%.
- Engineered Delta tables in Azure Databricks to support incremental loading, upserts, and optimized transformations using PySpark and SQL, efficiently processing 100GB of data.
- Implemented fact and dimension tables, SCD Type-2 logic, and automated Change Data Capture (CDC) solutions, optimizing data integration processes and reducing operational costs by 30%.
- **Tools:** Python, SQL, PySpark, Dataproc, Google BigQuery, Google Cloud Storage, Jenkins, Docker, Azure Data Factory, Azure Databricks

EDUCATION

Institute of Aeronautical Engineering

B.Tech. in Electrical and Electronics Engineering

Relevant Courses: Data Structures, Algorithms, DBMS, Machine Learning

Aug 2017 – Jun 2021

CGPA: 8.49/10

PROJECT WORK

- **Sentiment Analysis on Product Reviews:** Developed a machine learning model to classify product reviews as positive or negative using NLP techniques. Preprocessed text using tokenization, stopword removal, and TF-IDF. Built and compared Naive Bayes and LSTM models. *Tools: Python, scikit-learn, TensorFlow, NLTK; Dataset: Amazon Reviews (Kaggle)*
- **Real-Time Vote Counting and Winner Prediction System:** Built a streaming data pipeline using Apache Kafka and Apache Spark to process real-time vote data. Implemented dynamic winner prediction logic with fault tolerance and scalability. *Tools: Apache Kafka, Apache Spark Streaming, Python*

AWARDS AND CERTIFICATIONS

- **Google Cloud Certified:** Associate Cloud Engineer, Professional Data Engineer
- **Microsoft Certified:** Azure Data Engineer Associate
- Cleared Round 1 of the **TCS CodeVita 2020** global coding contest organized by Tata Consultancy Services