Arjun P

+91 96560 15203 | Chennai, Tamil Nadu, India | heyaitzarjun@gmail.com | linkedin.com/in/arjun−p | github.com/ArjHub | Portfolio

EXPERIENCE

Cloud Data Engineer, Tata Consultancy Services (TCS) Jul 2024 – Present

- Engineered ETL/ELT pipelines and Spark, Python and SQL workflows on AWS (S3, Glue, EMR, Athena, Redshift, Lambda, Step Functions) powering cloud data platforms across **30+ production deployments**.
- Implemented data quality and validation checks, including schema-aware processing and exception handling, improving consistency and reliability of downstream datasets.
- Built a real-time lakehouse platform on Databricks, PySpark, Kafka, Airflow and Delta Lake, processing **100K+ events** through Bronze/Silver/Gold layers with configurable validation rules and rejected-record tracking.
- Developed PySpark and SQL transformations for cleansing, enrichment, incremental processing and partitioned storage, orchestrated with Airflow for scheduling, dependency management and repeatable runs.
- Developed a Python metadata ingestion pipeline using PostgreSQL and REST APIs to collect, transform, validate and organize metadata from enterprise documents.
- Built FastAPI REST services with JSON-based data exchange and PostgreSQL-backed persistence for structured data access and integration.
- Automated application and infrastructure deployment using Docker, Terraform, Git and CI/CD, **cutting deployment effort by 70%**.
- Resolved production data and infrastructure incidents using CloudWatch and AWS tooling, leading root-cause analysis with cross-functional teams to improve pipeline reliability.

Machine Learning Engineer Intern, SmartInternz March 2024 – May 2024

- Developed regression, classification and clustering models, with data preprocessing and feature engineering to improve training data quality.
- Evaluated and tuned models (hyperparameter tuning, accuracy and reliability testing) to improve performance.
- Integrated a custom Python ML model into a Flask web application and supported its deployment and scalability.

PROJECTS

PipelineSentinel – Self-Diagnosing Data Pipeline Platform (github)

- Built a pipeline-observability platform that streams logs, metrics and data-quality results from Airflow and Spark jobs through Kafka-compatible Redpanda into Delta Lake, with idempotent event landing and a replayable journal so no events are lost during broker outages.
- Implemented detectors for failed jobs, schema drift, late data and volume anomalies (same-weekday median/MAD), with root-cause linking that collapsed 55 raw findings into 26 primary incidents; all 25 injected faults were detected, with zero false positives on 5 clean runs.
- Designed a Claude-based diagnosis agent using read-only tools, schema-validated output and evidence-quote verification, delivering ranked root causes to Slack with a signed, allowlisted, audited one-click retry; benchmark top-1 category accuracy was **81.3%** (25 faults $\times$ 3 repeats) at **$0.0025 per diagnosis**. The benchmark exposed a tool gap hiding out-of-memory errors (0/9 correct); fixing it raised accuracy to **93.3%** (post-hoc, same faults, not held-out).
- On 10 faults I also timed by hand, diagnosis took a median of **175 s manually vs. 13 s with the agent** (6/10 vs. 8/10 correct; small sample, partially blinded).

SKILLS

Languages: Python, SQL, PySpark, Bash, Java

Data & Streaming: Spark, Databricks, Delta Lake, Kafka, Airflow, PostgreSQL, Redshift

Cloud & DevOps: AWS (S3, Glue, EMR, Lambda, Athena), Terraform, Docker, CI/CD

APIs & ML: FastAPI, Flask, Scikit-learn, PyTorch

EDUCATION

B.Tech in Computer Science & Engineering, Specialization in Artificial Intelligence & Machine Learning 2020 – 2024
Vellore Institute of Technology, Chennai

ACHIEVEMENTS & CERTIFICATIONS

- Winner, TCS – Best Team Award, recognised as an inspiring role model by colleagues
- Winner, TCS – On the Spot (Team) Award for outstanding contribution to the organisation
- Winner, TCS – Applause for Team Award (2x) for outstanding contribution to the organisation
- Outstanding Performance Award – Tata Consultancy Services (TCS)
- AWS Certified Cloud Practitioner (CLF-C02)