Arijit Kumar Sahu

+91 7809222909 | arijitkumar2912@gmail.com | linkedin.com/in/arijitk1 | github.com/ProfessorXcommunity | Bangalore, India

Professional Summary

Data Engineer with 1.5+ years at TCS building cloud-native ETL pipelines, real-time streaming systems, and anomaly detection dashboards. Skilled in Spark, MongoDB, PostgreSQL, and AWS/GCP. Strong background in time-series analysis and scalable data architectures. Currently pursuing a PG Diploma in Data Science & AI at IIITB.

Education

Post Graduate Diploma in Data Science and Artificial Intelligence — Expected 2026
International Institute of Information Technology Bangalore (IIITB) — *Bangalore, India*

B.Tech. in Electronics and Telecommunications — May 2023
Indira Gandhi Institute of Technology — *Sarang, Odisha*

- CGPA: 8.56/10

Technical Skills

Languages: Python, SQL, Java, JavaScript
Big Data & Streaming: Apache Spark, PySpark, Airflow, Hadoop, Kafka, Data Warehousing & Lake
Cloud Platforms: AWS (S3, Glue, Redshift, EMR, Athena, Lambda), GCP, Databricks
Databases: MongoDB, PostgreSQL, MySQL, Cassandra, Redis
Analytics & Visualization: Tableau, Power BI, Atlas Charts, Pandas, NumPy, Matplotlib, Seaborn
DevOps & Tools: Docker, FastAPI, Git, Linux, CI/CD

Experience

Assistant System Engineer — Feb 2024 – Present
Tata Consultancy Services (TCS) — *Bangalore, India*

- **Data Engineer (Cisco Client):** Developing MongoDB Atlas fault-monitoring dashboards for a Cisco product tracking 36K+ devices and 25M+ daily events.
- Coordinating with cross-functional teams to design unified fault visibility dashboards.
- Performing time-series EDA and anomaly detection using PySpark & Pandas.
- Building Airflow ETL pipelines ingesting CSV/JSON/Parquet into MongoDB & S3.
- Optimizing MongoDB schemas & indexes, cutting query latency by 35%.
- **Database Administrator (Cisco Client):** Managing distributed MongoDB, PostgreSQL, Cassandra, and MySQL clusters supporting high-availability 70+ prod servers across multiple regions.
- Applied indexing and materialized views, cutting query time by 60%.
- Built backup & recovery (RPO 15m, RTO 2h) with Ops Manager & pg_dump.
- Automated schema migrations & monitoring with Python, reducing manual work by 40%.

Projects

Scalable Banking Data ETL Pipeline — Github — July 2025

- Designed PySpark ETL pipeline with multi-env configs; supported CSV & Hive ingestion.
- Streamed nested JSON events via Kafka for real-time processing.
- Built unit/integration tests (pytest, Chispa) achieving 95%+ coverage.
- Optimized Spark jobs with custom logging and 1000 shuffle partitions.

AWS Data Lake Architecture — Sep 2025

- Built AWS data lake using S3, Glue, Lambda, and Athena.
- Created landing, cleansed, and analytics zones for data flow.
- Automated ETL with Step Functions and Glue jobs.
- Connected Athena for query access and dashboards via QuickSight.
- Added CloudWatch alerts for pipeline monitoring.

Certifications

PySpark - Apache Spark Programming, Udemy (2025)

Management Consulting Essential Training, Udemy (2024)