Saumya Ranjan

saumya-ranjan/ ✉ saumyaranjan1111@gmail.com ☎ +91 8789404882

Education

Education	Institute	CGPA/%	Year
B.Tech	Birla Institute of Technology	9.1	2020–2024

Experience

Data Engineer, Axis Bank **Jul 2024 - Present**

Bengaluru, India

- Designed and maintained **40+** reusable, metadata-driven **ETL pipelines** using **IICS, Pyspark, Informatica** and **SQL** for financial data ingestion and transformation, ensuring data consistency, lineage, and governance compliance.
- Engineered data models to process **500K+ loan payments** and **1M+ credit card transactions** daily, identifying fraud and anomalies to feed risk mitigation systems.
- Optimized complex SQL queries in **Oracle SQL, Hive** and **Impala**, reduced execution time by **20-50%**.
- Integrated **SCD2** on fact table with **50M+** records, to ensure historical integrity and maintain robust data lineage.
- Collaborated with cross-functional business teams to build backend data models powering real-time **Tableau** dashboards.
- Automated **data quality checks** and report generation, saving **25+** hrs/mo of manual effort and improving accuracy.
- Implemented **CI/CD** pipelines with Jenkins and Bitbucket to streamline ETL deployments and ensure reliable releases.

Projects

Data Lakehouse Engineering: Incremental Customer Data Sync **PySpark, Iceberg, RDS, S3, Athena**

- Engineered an incremental batch ETL pipeline to sync Customer Profile data from an **AWS RDS** (PostgreSQL) source to an S3 Data Lake.
- Developed a PySpark job (hosted on EC2) to connect to RDS, extract new/updated records, and perform data transformations.
- Implemented Apache Iceberg to manage data lake table integrity, leveraging **ACID-compliant upserts** (MERGE INTO) to ensure reliable data consistency on update.
- Managed Iceberg metadata via the AWS Glue Data Catalog, enabling low-cost, serverless querying of the final dataset via **AWS Athena**.

Real-Time Log Aggregation with Stateful Streaming **PySpark, Kafka, Avro, Schema Registry, PostgreSQL**

- Engineered an E2E streaming pipeline using **Docker Compose** to orchestrate **Kafka, Schema Registry**, and PostgreSQL environments locally.
- Developed a data producer to send schema-enforced log events using **Avro serialization** and managed schema evolution via the Schema Registry.
- Built a **PySpark Structured Streaming** application to consume Kafka, implement stateful aggregation (1-min tumbling window) on log errors.
- Utilized a custom **foreachBatch sink** to reliably write time-windowed aggregates into a **PostgreSQL** table, ensuring sink flexibility.

ML Data Preprocessing for Speech Analysis **Python, Scikit-learn, Pandas, Librosa**

- Developed a robust **data processing workflow** to clean, normalize, and transform raw audio signals from a public dementia dataset.
- Engineered features by extracting **MFCCs (Mel-Frequency Cepstral Coefficients)** and other audio attributes using **Librosa** and **Pandas** to create a feature-rich dataset.
- Built and validated an SVM classification model (**Scikit-learn**) on the processed data, demonstrating an end-to-end data pipeline from raw unstructured data to a usable ML model.

Technical Skills

Programming & Query: Python — PySpark — Apache Spark — SQL — Shell (basics)
Cloud Services: AWS S3 — Lambda — RDS — Athena — AWS EC2
Data Lake & DBs: Apache Iceberg — Hive — Impala — Oracle SQL — PostgreSQL
Streaming & Ecosystem: Apache Kafka — Schema Registry — Avro — Informatica (IICS, PowerCenter)
DevOps & Tools: Docker — Jenkins — Bitbucket — Git
ML & Analysis: Pandas — Scikit-learn — Librosa

Achievements

- Received the **G.P. BIRLA Scholarship** for Academic Excellence (worth 1 lakh) for **4 semesters in a row**.
- **Specialist on Codeforces**, under 1000 AIR in multiple Global competitive programming contests, solved 500+ questions on Codeforces
- **Knight on Leetcode**, under 1000 AIR in multiple Leetcode weekly contests, solved 600+ questions on Leetcode