

Somya Pathak

☎ +1 720 936 7780 | ✉ somyapathak2910@gmail.com | 🔗 linkedin.com/in/somyapathak/

EDUCATION

University of Colorado, Boulder

Master of Science in Computer Science **GPA: 4.0/4.0**

Boulder, CO, United States

Aug 2024 – May 2026

MKSSS Cummins College of Engineering for Women - Pune University

Bachelor of Technology in Computer Science; **GPA: 3.56/4.00**

Pune, Maharashtra, India

Aug 2017 – Jul 2021

SKILLS

Technologies: Machine Learning, Hadoop, HDFS, Hive, Kafka, Apache Spark, Amazon S3, CEPH, MLflow, Kubeflow, Apache NiFi, Splunk

Languages: Python, Java, Scala, Groovy, SQL, Spark

Databases: MySQL, Snowflake, DBMS, Neo4j, Redis

Tools: Eclipse, IntelliJ, VS Code, Git, Jenkins, Rally, SQL Developer, Toad, Jupyter Notebook, Apache NiFi, Splunk, SFTP

WORK EXPERIENCE

Mastercard Technologies - Data Engineer

Jul 2021 – Jul 2024

Pune, Maharashtra, India

- **Real-Time Data Ingestion and Processing Framework using Spark & Scala:**
 - * Designed, developed and deployed a real-time data pipeline using Apache Spark Scala and Apache NiFi, orchestrating high-throughput Kafka stream ingestion and transformation, processing over **10M events/day**.
 - * Parsed raw events into structured, masked formats for real-time loading into the Data Warehouse with **100% uptime**, storing outputs in S3-MinIO and CEPH across multiple availability zones for fault tolerance.
 - * Collaborated across Agile sprints, participating in sprint planning, daily Scrum calls, and retrospectives to ensure timely and iterative delivery.
- **Data Regularization & Visualization:**
 - * Architected and deployed Apache NiFi pipelines and SQL-based data loaders, streamlining the ingestion and transformation of structured datasets across multiple sources.
 - * Integrated Splunk dashboards & alerts for real-time monitoring, enabling proactive detection of anomalies and performance issues while ensuring data integrity through automated validation and logging.
- **Real-Time Reconciliation & Auditing Framework with Spark and Kafka (Axon):**
 - * Designed and developed a reconciliation utility to track and validate streaming Kafka messages from Axon, ensuring **100% data accuracy** across the Data Warehouse.
 - * Engineered Spark-Scala-based auditing to perform high-throughput data validation, consistency checks across AZs, and audit logging; published verified results back to Axon.
 - * Built reusable modules with rigorous unit testing and CI/CD integration; maintained production stability through real-time Splunk dashboards and automated anomaly alerts.
- **Data Retention & Purging Utility for Storage Optimization:**
 - * Designed a configurable purging utility with automated validation, to enforce data retention policies, improving storage efficiency and system performance with **0%** impact on active datasets.
 - * Utilized Python and Jupyter Notebooks for data analysis, trend visualization, and cleanup cycle optimization.
- **Data Transformation & Transfer Solutions:**
 - * Delivered scalable data transformation and transfer solutions, integrated with client APIs over secure VPN connections.
 - * Optimized Groovy scripts for vault connectivity and developed high-performance Spark-Scala pipelines to fetch data from HDFS.
 - * Enabled secure data delivery via SFTP for customer processing within **24** hours of transaction and streamlined workflows for automation and scalability.
 - * Recognized by stakeholders for innovative, timely solutions that enhanced project reliability and delivery.

IBM Technologies - Cloud Integration Intern (IBM Security Verify)

Sep 2020 – Jun 2021

Pune, Maharashtra, India

- **PaaS-based Application for Security & Access Management:**
 - * Engineered and deployed a cloud-based React helpdesk platform by integrating multiple IBM APIs, including IBM Security Verify, to enhance authentication, access management, and overall security workflows.
 - * Implemented OIDC-based authentication to securely generate access tokens for users and tenant administrators.

Mastercard Technologies - AI/ML Intern

May 2020 – Jun 2020

Pune, Maharashtra, India

- **Prediction Model for Performance Evaluation:**
 - * Developed a predictive analytics pipeline using **9 machine learning models**, achieving a highest accuracy of **93%** across diverse datasets, while gaining end-to-end experience in model inception, training, optimization, and evaluation workflows.
 - * Performed comparative analysis of MLflow and Kubeflow for model tracking and lifecycle management, emphasizing scalability and reproducibility in MLOps.

AWARDS & ACHIEVEMENTS

- Won the **Goldman-Sachs We-Tech Scholarship and Mentorship** for outstanding fellow in ICT field, 2019.
- **Runner up** in IBM Hackathon on Climate Change, 2020.