Gaurav Singh Bisht

LinkedIn | gauravsinghbisht.scoe.it@gmail.com | 9510158812 | Github | Pune, India | DOB: 26-03-2004

Summary

Aspiring Data Engineer with a strong foundation in distributed data systems, cloud platforms, and real-time analytics. Skilled in building scalable ETL pipelines, developing Spark-based data applications, and integrating modern cloud-native architectures. Proven experience in both academic and industry-led projects involving big data, data warehousing, and machine learning.

Education

Savitribai Phule Pune University, BE in Information Technology — Sept 2021 – June 2025

- CGPA: 8.4/10.0

Technical Skills

Languages: Java, Python, SQL

Big Data Tools: Apache Spark, Apache Kafka, Amazon S3, Snowflake, ETL, HDFS, AWS Glue, AWS Athena

Technologies: TensorFlow, Nodejs, Reactjs

Tools: Git, Docker, MongoDB, MySQL, Databricks

Experience

Big Data Intern, Accenture – Pune, India — Feb 2025 – June 2025

- Worked in Supply chain department on a pharmaceutical client project within the Data Integration Team, focusing on end-to-end data flow design and pipeline integration between SAP and Kinaxis.
- Collaborated with cross-functional teams to understand domain-specific data, ensuring accurate mapping, transformation, and delivery of datasets.
- Gained hands-on experience with enterprise data systems, enhancing understanding of large-scale data ingestion and integration best practices.

Backend Intern(Nodejs), CNear – Delhi, India — Dec 2023 – March 2024

- Developed a centralized CRUD API system, improving backend performance by 80%.
- Integrated RBAC using JWT to secure over 20 endpoints. Handled 1,000+ API calls weekly.

Projects

LogLoader - HDFS Log Parser Pipeline using Spark (Java)) — Github Link

- Designed and implemented a scalable Apache Spark application in Java to parse 11M+ raw log records into structured templates for downstream analytics.
- Leveraged Spark optimizations like caching and broadcasting to reduce batch processing time to 5 minutes.
- Focused on generating clean and structured log datasets for anomaly detection, monitoring, and visualization.

STEDI Human Balance Analytics (Udacity Capstone) — Github Link

- Developed a data lakehouse pipeline on AWS to curate mobile and sensor data from a human balance trainer device.
- Created ETL workflows in Glue to transform raw accelerometer and step sensor data. Designed Athena-compatible curated tables used to train a real-time step detection ML model while maintaining strict privacy compliance.

Snowflake Restaurant Ratings datawarehouse project (Udacity Capstone) — Github Link

- Built a cloud-native OLAP data warehouse using Snowflake, integrating Yelp reviews with GHCN-D weather data to analyze correlations between ratings and weather. Designed staging, ODS, and star schemas; transformed nested JSONs; and delivered insights on temperature, precipitation, and review scores.

Deepfake Detective - GAN-Based Deepfake Detection Research) — Github Link

- Developed a deepfake detection model using GAN discriminators, achieving 99% accuracy and 0.005 loss. Demonstrated the potential of GAN discriminators as classifiers while revealing their limited generalizability across GAN architectures.

Achievements and Certifications

- Rockwell Automation Hackathon Winner: Led a team in developing an AI-powered solution to win a national-level hackathon. - Project Link
- Big Data and Hadoop for ATCI Stream - Certificate Link