Prathik Bharath Jain
Denver, CO, 720-990-4940 | prathikbafna0@gmail.com| LinkedIn | GitHub

SUMMARY

Data Engineer skilled in building scalable ETL/ELT pipelines, streaming systems, and MLOps workflows using Azure, Spark, Airflow, Databricks, and Snowflake. Experienced delivering production data systems across manufacturing, utilities, media.

EDUCATION

Master of Science in Data Science - University of Colorado Boulder, USA CGPA: 3.9/4.0 **August 2023 - May 2025**

Bachelor of Technology in Computer Science - PES University, India CGPA: 8.4/10 **August 2018 - June 2022**

TECHNICAL SKILLS

Languages: Python, R, Java, Scala, SQL, C, JavaScript

Data Engineering: Azure Data Factory, Azure Synapse, Databricks, Spark, Airflow, Kafka, MongoDB, Hive, Hadoop, MySQL
Cloud & Infra: Azure, AWS (S3, Glue, Athena, SageMaker, lambda, Redshift), GCP(BigQuery), Docker, Kubernetes, Git
Analytics & BI: Power BI, Tableau, Snowflake.

EXPERIENCE

Data Scientist | Atkins Realis | Denver, CO, USA — July 2025 - Present

- Developing an ML model to predict water pipeline bursts using feature engineering to reduce failures and disruptions.
- Built a Power BI dashboard to monitor grant applications across pre- and post-award workflows, integrating multiple data sources and automating refresh schedules.

Data Engineering Intern | Lincoln Electric | Cleveland, OH, USA — May 2024 - May 2025

- Built 75+ data pipelines in **Azure Data Factory** using Azure data bricks, and medallion architecture to transform and migrate data for **Power BI**, facilitating business intelligence and creating **SQL scripts** for ad-hoc requests.
- Extracting tables from **SAP** tables using OData connectors and SAP table connectors to the Azure Synapse environment building the **Data warehouse** environment and scheduling update pipelines in Azure Synapse Analytics.
- Transcribed Oracle **stored procedures** and scripts into SQL Server and migrated data flows from **Alteryx** to Azure.
- Developed a comprehensive dashboard for tracking log metrics in real-time; this initiative empowered teams with actionable insights by reducing troubleshooting time by 60% during critical operational periods.

Data Scientist | Honda 99P Labs | Boulder, CO, USA — January 2024 - May 2024

- Ingested and processed V2X telematics data (61,930 trips, 157 drivers) into analysis-ready tables using Python and SQL, then built driver profiling models to evaluate speeding, braking, and lane-changing behaviors for road safety analytics.

Machine Learning Engineer | Gil's Lab | Boulder, CO, USA — September 2023 - May 2024

- Built an end-to-end computer vision pipeline using YOLOv8 and Bot-SORT for small-object fish detection and tracking, achieving **93.87%** accuracy and enabling automated behavioral analytics.

Software Engineer | Zee Entertainment Ltd | Bangalore, India — January 2022 - July 2023

- Engineered **Java Spring Boot** framework to develop the new Juspay Payment page integration. Ensured seamless compatibility with the existing infrastructure and successfully transitioned **95%** of total payment traffic to the platform.
- Revamped **Apple payment** token verification process to in-house verification and reduced the API time by **92%**.
- Implemented Asynchronous Kafka communication between microservices that increased transaction success rates.

PROJECT

SQL RAG - LLM Powered SQL Query | Python, LLM, RAG, Streamlit, SQLite, REST API

- Developed a Streamlit web app that converts natural language into SQL queries using a local LLM with Retrieval-Augmented Generation (RAG) over schema metadata, improving query accuracy on complex joins.
- Automated table creation, data insertion, and real-time query execution on SQLite, streamlining data retrieval and manipulation.

Sales Forecasting MLOps Platform | Python, Airflow, MLflow, Docker, Minio

- Built a **production-grade MLOps pipeline** using Astronomer/Airflow to orchestrate data validation, feature engineering, model training, experiment tracking, and artifact storage for forecasting.
- Implemented an ensemble forecasting system (XGBoost, Prophet) with Optuna optimization, achieving low MAPE, comprehensive diagnostics through automated visualization and MLflow-tracked metrics.

Stream Forge - Real-time Data Streaming | Python, Airflow, Spark, Kafka, Databases, AWS

- Implemented a real-time data streaming pipeline using Apache Airflow to fetch the information from an API, pass it to Kafka stream, and store it in PostgreSQL, which is further transformed by spark and stored in Apache Cassandra.
- Developed a pipeline to extract data from an AWS S3 bucket into AWS Glue Data catalog using Glue Crawlers and created an ETL job to transform the data using Spark, Lambda, and stored in the Parquet format and analyzed using Athena.