Ajay Raj Singh

Email: ajayrajsingh2003@gmail.com, **Address**: Jersey City, NJ | +1 (732)-209-0281
LinkedIn: www.linkedin.com/in/connectwithajayrajsingh, GitHub

Summary

Data Engineer with 6+ years of experience delivering production-ready data solutions and scalable ETL pipelines. I am expert in SQL, Python, and data visualization tools, with a strong track record of optimizing workflows in cloud environments. Demonstrated ability to collaborate with teams to maintain data quality and support analytic product development through effective documentation and project management.

Skills

- **Programming Languages & Databases:** Python, SQL, PL/SQL, PostgreSQL, MongoDB, Snowflake, Redshift, RDS
- **Big Data & Data Engineering:** PySpark, Apache Spark, Apache Airflow, ETL/ELT Development, dbt, Data Warehousing, Data Modeling (Star Schema), Data Partitioning, Query Optimization, REST API Integration, JSON Data Processing
- **Cloud Services, DevOps & Infrastructure:** AWS (S3, EMR, Glue, Athena, Lambda, EC2, EKS, CloudWatch, CloudFront, API Gateway, S3 Event-Driven Pipelines), Azure (Data Factory, Maps API, Databricks), Docker, Kubernetes (EKS), Jenkins, Git, GitLab, CI/CD Pipelines, Infrastructure as Code, SaaS, PyODBC
- **Data Quality & Monitoring:** AWS CloudWatch, Data Validation, Schema Drift Detection, Automated Alerting, Data Integrity Checks, SLA Monitoring
- **Analytics & Visualization:** Pandas, Pytorch, NumPy, Scikit-learn, Plotly, NLP (Text Processing, Sentiment Analysis), Feature Engineering, Machine Learning, Flask APIs

Experience

Pavane Solutions Inc. ***Data Engineer*** *Remote, NJ* | **Jul 2024 – Present**

- Engineered production PySpark ETL pipelines on AWS EMR/Glue processing 200GB+ daily healthcare claims and eligibility data into S3/Redshift/Athena with 99.5% uptime.
- Architected Apache Airflow DAGs orchestrating 50+ data workflows across AWS services, implementing retry logic and SLA monitoring to achieve 40% reduction in manual intervention.
- Migrated legacy on-premises Oracle database to Snowflake cloud data warehouse, designing star schema models and implementing incremental ELT using dbt for dimensional analytics.
- Optimized PySpark jobs through broadcast joins and dynamic partitioning strategies, reducing processing time by 35% and cutting EMR costs by $1,800/month on multi-TB datasets.
- Implemented comprehensive data quality framework using AWS CloudWatch metrics and Python validators, creating automated alerts for schema drift, null checks, and referential integrity violations.
- Deployed containerized data applications using Docker on AWS EKS, establishing CI/CD pipelines with Jenkins and Git for automated testing and deployment of ETL code.

Saint Peter's University ***Data Science Researcher*** *Jersey City, NJ* | **Nov 2023 – Feb 2025**

- Engineered scalable NLP data pipeline processing 10K+ daily news articles from enterprise REST APIs, implementing PySpark transformations for text cleaning and sentiment analysis workflows.
- Built real-time geospatial ETL system integrating Azure Maps API and Air Quality Index data using Python and Azure Data Factory, optimizing healthcare route planning for 200+ facilities.
- Developed PySpark-based healthcare data pipeline for breast cancer diagnostic analytics, implementing feature engineering and data preprocessing achieving 15% model accuracy improvement.
- Deployed web analytics platform on AWS using Lambda and S3, reducing stakeholder content delivery latency by 65% through CloudFront CDN integration and optimized query patterns.
- Presented 2 research papers on data engineering best practices at NJBDA conference, focusing on real-time pipeline architectures and cloud-native ETL optimization strategies.

IT Nopal Technologies ***Data Scientist II*** *New Delhi, India* | **Jun 2021 – Jan 2023**

- Designed and maintained high-throughput data pipelines for sports analytics platform processing 50K+ match events weekly, integrating REST APIs with SQL-based ETL systems using Python and Airflow.
- Optimized AWS infrastructure reducing cloud costs by 15% through right-sizing EC2 instances, implementing S3 lifecycle policies, and refactoring Lambda functions for efficient memory usage.
- Collaborated with DevOps teams to containerize ML model training workflows using Docker, establishing GitLab CI/CD pipelines for automated testing and deployment to production environments.
- Automated SQL data quality processes using Python scripts, implementing scheduled validation checks that improved data integrity to 100% across PostgreSQL databases and reduced manual QA time by 20%.
- Built real-time data ingestion pipeline using AWS Lambda triggered by S3 events, processing JSON payloads from sports APIs and loading structured data into RDS with 30% faster retrieval.

IT Nopal Technologies ***Data Scientist I*** *New Delhi, India* | **Jan 2019 – May 2021**

- Led database migration from NoSQL (MongoDB) to PostgreSQL using Python PyODBC, redesigning schema structure that reduced average query execution time by 85% and EC2 compute costs by 70%.
- Developed serverless data access layer using AWS Lambda and API Gateway, enabling real-time retrieval of customer analytics data with 30% performance improvement over legacy EC2-based architecture.
- Created interactive business intelligence dashboards using Plotly and SQL queries against PostgreSQL, visualizing product KPIs and operational metrics for executive stakeholders.

- Built interpretable ML models for classification and prediction using Pytorch, deployed as part of customer analytics workflows with 82% accuracy for business stakeholders.

Projects

Interactive Keyword-Based News Retrieval System – NLP & Real Time Content | **NJBDA** – *Under Review (*Link**,** GitHub**,** Website)

- Developed a real-time NLP-powered news retrieval platform integrating REST APIs (NYT, Bing) and Streamlit, automating content delivery and boosting user engagement through interactive keyword driven visualizations.

Optimization with Metaheuristic Algorithms – Breast Cancer Diagnosis | *Saint Peter's University Symposium (*GitHub)

- Implemented a hybrid Ant Colony Optimization and Grid Search framework, enhancing model accuracy by 15% and cutting computation time by 20% for efficient, data-driven diagnostic predictions.

Health-Centric Navigation & Air Quality Management for Sensitive Populations (Map routing) | **NJBDA** – *Under Review (Link*, GitHub, Website)

- Engineered least polluted routing algorithm using Azure Maps and live AQI data to minimize pollution exposure, validating performance across world's most polluted cities. (Mock Data for Website)

Achievement & Honors

- **Data Science Club President** (*Sep 2024 – Apr 2025*) | Led 50+ member organization *(Link*)
- **Lead Presenter – NJBDA** Research Conference & Symposium 2025 |Presented **2 Research Papers** on Data Engineering/Science *(Link*)
- **1st Prize** – Data Science Showcase | Real-time least polluted routing application *(Link*, GitHub)
- **Alpha Sigma Nu** Honor Society Inductee | National Jesuit honor society recognizing academic excellence *(Link*)
- **Data Storyteller** Award Winner | Recognized for data visualization and communication excellence *(Link*, GitHub)

Education

Saint Peter's University, Master of Science in **Data Science** (**GPA - 3.95**/4.00) Jersey City, NJ| **Feb 2023 – Feb 2025**