

Sajal Bhattarai

Data Engineer | Hicksville, NY

516-939-8077 | jobsajalbhattarai@gmail.com | [Linkedin](#)

SUMMARY

Results-driven **Data Engineer** with around **5 years** of experience delivering scalable, production-grade data platforms across cloud environments. Expertise in **Python, SQL, Apache Spark, and Airflow**, with hands-on experience building **ETL/ELT pipelines** and data warehouses on **AWS, Azure, and GCP**. Proven track record of implementing dimensional data models and reliable data systems that enable analytics, AI-driven insights, and data-informed decision-making through cross-functional collaboration.

TECHNICAL SKILLS

Languages	Python, SQL, Bash
Big Data & Processing	Apache Spark, Databricks
Data Engineering	ETL / ELT Pipelines, Batch Processing, Incremental Loads, Data Validation, Data Quality Frameworks
Data Modeling & Warehousing	Dimensional Modeling, Star Schema, Fact & Dimension Tables, Data Warehousing
Databases & Platforms	PostgreSQL, MySQL, Snowflake, Amazon Redshift, Google BigQuery, Vector Databases (pgvector)
Cloud Platforms	AWS (S3, EC2 concepts), Microsoft Azure, Google Cloud Platform (BigQuery, Cloud Storage)
Orchestration & Streaming	Apache Airflow, Apache Kafka
DevOps & Tools	Git, Docker, Linux, CI/CD Concepts

EDUCATION

Bachelor’s in Computer Science | State University of New York, Old Westbury | Old Westbury, NY

EXPERIENCE

Data Engineer | KWI - Melville, NY Nov 2024 – Present

Responsibilities:

- Designed and implemented **production-grade ETL/ELT pipelines** ingesting data from MySQL and internal systems into PostgreSQL and Snowflake-based data warehouses.
- Built **Airflow DAGs** to orchestrate batch pipelines with dependency management, retries, and monitoring.
- Developed **Spark-based transformation jobs** using Databricks to process large datasets efficiently and improve pipeline performance.
- Designed **dimensional data models (fact and dimension tables)** to support product analytics, operational reporting, and downstream BI tools.
- Engineered an **AI-powered semantic search platform** using RAG architecture and PostgreSQL pgvector, enabling similarity search across 1,700+ records.
- Integrated **Azure OpenAI** into data pipelines to enrich and classify product feedback data, reducing manual analysis by 3–5 hours per week.
- Implemented **data quality checks** to validate schema integrity, record completeness, and consistency before production releases.
- Optimized SQL queries and indexing strategies to improve analytics query performance.
- Collaborated with engineering, QA, and product teams using **Git-based workflows** and Agile practices.
- Authored detailed technical documentation covering pipeline architecture, data models, and operational runbooks.

Environment: Python, SQL, Apache Spark, Databricks, Apache Airflow, PostgreSQL, MySQL, Snowflake, Azure, Azure OpenAI, Docker, Linux, Git

Teaching Assistant | State University of NY – Old Westbury, NY

Aug 2023 – Oct 2024

Responsibilities:

- Supported graduate-level Data Engineering courses focused on ETL pipelines, data warehousing, and distributed data processing.
- Guided students in building Spark-based batch data pipelines using Python and SQL.
- Taught dimensional modeling techniques for analytical workloads and reporting systems.
- Demonstrated cloud-native data architectures using GCP BigQuery and Cloud Storage, alongside AWS and Azure.
- Evaluated projects involving Airflow orchestration, Spark transformations, and scalable analytics pipelines.
- Mentored students on Linux, Git, debugging, and production readiness for data systems.

Environment: Python, SQL, Apache Spark, BigQuery, Cloud Storage (GCP), PostgreSQL, MySQL, Airflow, Linux, Git

Data Engineer | Verisk Analytics | Jersey City, NJ

Jan 2019 – May 2021

Responsibilities:

- Built and maintained **enterprise ETL pipelines** ingesting high-volume data into centralized data warehouses.
- Developed **ELT workflows** using Python and SQL in Amazon Redshift environments.
- Implemented **Spark batch processing jobs** to transform large datasets efficiently.
- Designed and optimized **data warehouse schemas** to improve query performance and reporting speed.
- Orchestrated pipelines using **Apache Airflow**, ensuring reliability and observability.
- Worked with **Kafka-based streaming pipelines** for near real-time data ingestion.
- Containerized data jobs using **Docker** to standardize deployments.
- Automated operational tasks using **Linux and Bash scripting**.
- Collaborated with analytics and data science teams in Agile sprints.
- Maintained technical documentation for pipelines and data models.

Environment: Python, SQL, Apache Spark, Amazon Redshift, Apache Airflow, Apache Kafka, Docker, Linux, Bash, Git, AWS