

RUPA BHATIA

New York | rupa.bhatia@myjobsmails.com | +15168207590 | [rupa-portfolio](#) | [Linkedin/rupa-bhatia](#) | [github/Rupa2400](#)

EDUCATION

Syracuse University, Syracuse, NY

May 2024

Master of Science in Computer Science

University of Mumbai, Mumbai, India

May 2022

Bachelor of Engineering in Computer Engineering

EXPERIENCE

Morgan Stanley | Data Engineer

Jul 24 - Current

- Engineered a **cloud-native** data **ingestion** framework using **Apache Kafka**, **AWS Glue**, **S3**, and **Spark**, unifying **real-time** and **batch data** sources across teams, improving **data availability** by **30%**, and enabling scalable ingestion for **tick data** and regulatory **inputs**.
- Automated processing of **0.5 TB+** financial data using **Python**, **Spark**, **SQL** and **Redshift**; integrated **Delta Lake** on **AWS S3** with **ACID compliance** and time-travel features, improving data reliability, **pipeline throughput**, and report **delivery times**.
- Built modular, metadata-driven **ETL** pipelines with **Apache Airflow (DAGs)**, reducing **daily job failures** by **30%** and cutting **failure recovery** times by **60%**, while ensuring **regulatory compliance (GDPR, CCPA)**.
- Executed a **petabyte-scale migration** from **Hadoop** to **Snowflake**, moving **1.2 TB+** of **tick-level trade data** via **AWS S3**, increasing analytical throughput by **25%** and cutting on maintenance costs by **30%**.
- Standardized **40+ transformation models** using **dbt**, implementing testable, reusable **SQL** patterns across **10+ financial datasets**, reducing redundancy and improving data reliability across financial analytics workflow.

Intuit | Business Intelligence Intern

Jan 2024 - Jun 2024

- Engineered **end-to-end** data pipelines leveraging **Azure Databricks (Python, Pandas)** and **SQL** to optimize **inventory integration** and management for **50,000+ SKUs** across multiple **warehouses**, reducing **stockouts** by **18%** and lowering inventory **holding costs** by **12%**.
- Automated complex transportation route optimization by designing scalable data pipelines with **Azure Data Factory** and **Azure Functions**, improving delivery efficiency by **17%** and minimizing **bottlenecks** in high-demand distribution centers.
- Created interactive **Power BI** dashboards integrated with **SQL Server Analysis Services (SSAS)** and **Delta Lake**, enabling real-time visualization of operational and financial **KPIs**, increasing actionable insights by **22%** and enhancing **decision-making** across teams.

Infinite Infolab | Senior Data Engineer

May 2021 - Jul 2022

- Architected and implemented a cloud-native **AWS** data platform (**S3**, **Redshift**, **Glue**, **Lambda**) reducing storage costs by **20%** and boosting query performance by **25%**, enabling faster and more cost-effective **financial mortgage** data analysis.
- Automated data quality monitoring and event-driven workflows with **AWS Lambda**, integrating **OLAP** cube processing cutting manual data validation efforts by **1,000+** man-hours annually and reducing data quality issues by **30%**.
- Built interactive dashboards in **AWS QuickSight** and **Tableau**, in driving a **20% increase** in data-driven decisions by visualizing key portfolio **KPIs** and performance metrics.

Vivma Software Inc | Data Engineer

Feb 2020 - Apr 2021

- Streamlined **ETL** workflows using **Python**, **SQL**, and **SSIS**, improving data accuracy by **20%** and reducing manual efforts by **30%**.
- Built **Power BI** dashboards and **Excel** tools integrating **RDBMS** data, cutting decision time from **2 days** to **12 hours** and boosting inventory turnover **1.5x**.
- Implemented **CI/CD** pipelines with **Jenkins** and **Git**, cutting deployment errors by **40%** and speeding releases by **25%**, while monitoring **KPIs** to reduce production bottlenecks by **30%**.

TECHNICAL SKILLS

- Programming & Tools:** Python, R, SQL, NumPy, Pandas, OpenAI, JavaScript, PySpark
- Big Data & Cloud:** AWS (EC2, S3, Glue, Athena, EMR, Redshift, Lambda), Azure (Data Factory, Databricks, Delta Lake, Data Lake Storage, SQL Database), Hadoop, MapReduce, Apache Airflow, ETL, CloudWatch, A/B Testing, Apache Kafka, Apache Flink
- Data Engineering:** Data Ingestion, Data Extraction, Data Transformation, Data Validation, Data Architecture, Data Governance, Metadata Management, Data Cataloging, Data Pipelining, Data Orchestration, Data Modeling, Data Integration
- Data Warehousing:** Snowflake, Azure Synapse Analytics, Schema Design, Query Optimization, Performance Tuning
- Data Science & Machine Learning:** Time Series Analysis, Deep Learning, Natural Language Processing (NLP), Model Deployment, Chatbots, Large Language Models (LLMs), Linear & Logistic Regression, Gradient Boosting, Decision Trees, Random Forest, TensorFlow, Keras, PyTorch, Scikit-Learn, SciPy, Collaborative Filtering
- Data Visualization:** PowerBI, Tableau, Matplotlib, MS Excel, AWS QuickSight, MicroStrategy, ReportWORQ
- DevOps Tools:** Jenkins, Git, Docker, GitHub Actions, Kubernetes, Jira, Confluence, CI/CD