Vinay Kiran Raju

Data Engineer

raj.vinay2408@gmail.com | +1 201-492-8306 | LinkedIn | Portfolio | Blog

PROFILE SUMMARY

Results-driven Data Engineer with 4+ years of experience designing, building, and optimizing scalable data pipelines, ETL/ELT solutions, and real-time analytics platforms. Highly skilled in the **Microsoft Azure ecosystem** including **Azure Data Factory (ADF), Azure Synapse Analytics, Azure Databrick** and **Amazon Web Services (AWS)** including **S3, Redshift, and Glue**. Proven expertise in implementing advanced techniques such as **event-driven architecture, incremental data loading, metadata control tables, and PySpark** for high-volume data. Adept at leveraging **Python** and **SQL** to drive predictive modeling, root-cause analysis, and dashboard reporting that enhance business decision-making

TECHNICAL SKILLS

- **Programming:** Python (Pandas, NumPy, Scikit-Learn, Matplotlib, Plotly), PySpark SQL, Hadoop, Scala.
- **Data Engineering:** ETL/ELT, Data Modeling, Data Lakes, Incremental Loading, Metadata Management, Distributed Systems (PySpark, Spark SQL)
- **Cloud:** Azure (ADF, Synapse Analytics, Databricks, ADLS Gen2, Azure Functions, Unity Catalog), AWS (S3, Redshift, Glue, Lambda, DynamoDB)
- **Orchestration:** Azure Data Factory (ADF), Apache Airflow/Cloud Composer, AWS Glue
- **Databases:** MySQL, Azure Synapse Analytics, Snowflake, Redshift
- **Visualization & Tools**: Tableau, Matplotlib, Seaborn, Git, Microsoft Excel (VBA, Macros)
- **Best Practices:** Agile (Scrum), Git, Jira, CI/CD, Data Governance, Data Validation, Unit Testing

WORK EXPERIENCE

Cognizant **New Jersey, USA**

Data Engineer *Apr 2024 - Present*

- Engineered **scalable, event-driven data ingestion pipelines** using **Azure Functions** triggered by file placement in **Azure Data Lake Storage (ADLS Gen2)**, **optimizing the high-volume data workflow** for 10+ data sources and reducing data latency by **20%**.
- Designed and implemented a robust ELT architecture using **Azure Databricks (PySpark)** and **Delta Lake** principles, centralizing data assets with **Unity Catalog** for improved governance and data discovery.
- Orchestrated complex data workflows across the Azure ecosystem using **Azure Data Factory (ADF) pipelines**, which incorporated dynamic mapping, advanced error-handling, and automated scheduling for 24/7 reliability.
- Utilized **Azure Synapse Analytics** for highly efficient, large-scale data warehousing and analytical query execution, resulting in a **15% improvement in query performance** for reporting stakeholders.
- **Implemented robust incremental data loading strategies** using **ADF control tables** to track high-water marks and ensure non-redundant, efficient processing across daily pipeline runs.
- Applied stringent data quality control by implementing **schema validation** and automated data quality checks (DQCs) using **Databricks notebooks**, ensuring **99.5% data integrity** before consumption.

Natsoft Corporation **New Jersey, USA**

Data Engineer *Mar 2023 - Apr 2024*

- Developed **real-time payment analytics pipelines** on AWS, successfully processing the **transactional data** using **AWS Lambda** and **S3-based data lakes**.
- Optimized ETL loads by leveraging **AWS Glue** and **PySpark** for pre-processing, and utilized **Redshift staging tables** for complex, highly optimized transformations, specifically improving final query performance by **40%**.
- **Designed and enforced incremental load logic** within the S3 data lake structure, utilizing a **control table** to manage file ingestion state and prevent reprocessing of high-volume transactional data.
- Developed and deployed a robust **data validation framework** using **Apache Airflow DAGs** for end-to-end orchestration, cutting payment analytics reporting time by **30%** via automated data delivery to Tableau dashboards.
- Implemented **schema evolution** logic and compliance-driven data quality checks using PySpark, which significantly reduced data anomalies and increased reporting accuracy by **30%**.
- **Maintained and optimized the foundational data lake** by implementing **S3 lifecycle policies** and partitioning strategies, ensuring scalable and **cost-effective storage** for all ingested transactional data.

Hyderabad, India

Data Analyst *Sep 2020 - Jan 2022*

- **Processed and cleansed 5M+ records** using **Python** and advanced **SQL** to facilitate **root-cause analysis** of major operational events and identify key performance drivers.
- Created **real-time interactive Tableau dashboards** that centralized critical business data, reducing manual reporting cycles by **40%** for a user base of over 50 stakeholders.
- Performed **feature engineering** and data quality optimization, including handling missing values and encoding categorical data, to ensure model reliability and achieve high predictive performance.

- Built and validated **machine learning models** for **predictive modeling for high-value outcomes**, enhancing team decision-making efficiency by **25%** and leading to more precise strategic interventions.

PERSONAL PORTFOLIO PROJECTS

Zillow End-to-End Automated ETL Pipeline *Jan 2024 - Feb 2024*

- Developed an end-to-end ETL pipeline for **Zillow**, extracting real estate data from their API, transforming it on an **EMR** cluster using **PySpark**, and loading it into S3.
- Extracted Automated the workflow with **Apache Airflow** and integrated **Tableau** for data visualization, showcasing expertise in data transformation and analysis.

Loan Default Prediction *Jan 2023 - Mar 2023*

- Developed a Python-driven **Decision Tree ML model** for loan default prediction with **95.21% accuracy and 92.17% AUC** on 35,747 entries.
- Optimized Improved data quality by handling missing values, encoding categorical features, and eliminating duplicates; visualized key insights using **Tableau**, enhancing loan risk assessment.

EDUCATION

Pace University, Seidenberg School of Computer Science and Information Systems **New York, USA**
MS in Data Science | Concentration: Data Analysis, Visualization & Machine Learning *Date: May 2023*

Gandhi Institute of Technology and Management **Andhra Pradesh, India**
Bachelor of Technology in Electronics and Communication Engineering *Date: Apr 2021*

CERTIFICATES

- **Databricks Certified Associate Data Engineer –** Databricks, Nov 2025
- **Kore.ai Advanced Chatbot developer –** Kore.ai, Mar 2024
- **Snowflake BUILD 2023 LLM Bootcamp** – Snowflake, Dec 2023
- **Azure Data Fundamentals** – Microsoft, Apr 2023
- **Tableau Desktop Specialist** – Tableau, Mar 2023
- **Google Data Analytics Professional Certificate** – Coursera, Nov 2022