Vishnu Priya Neerukattu
E-mail: vishnuss8322@gmail.com
Phone: (940)-843-8066
LinkedIn: Vishnu Priya Neerukattu | LinkedIn

SUMMARY

- Experienced **Data Engineer** with 5+ years of success in designing and managing **data pipelines**, integrating **cloud platforms** (AWS, Azure), and delivering **analytics solutions** with tools like **SQL, Python,** and **Tableau**. AWS Certified Solutions Architect Associate with deep knowledge of **ETL processes**, **data modeling**, and **big data frameworks**. Adept at transforming complex datasets, including geospatial and healthcare data, into strategic insights in **agile, cross-functional** teams.Performed **data linkages between public health datasets** and **geospatial data assets** using tools like **GeoPandas**, **ArcGIS**, and spatial SQL, enabling location-based epidemiological insights.
- Documented and versioned all **ETL workflows** and **data transformation logic** to ensure full reproducibility, audit compliance, and transparency in data workflows.
- Tested datasets and applications to **validate data accuracy, completeness,** and **quality**, applying data validation techniques and automated test frameworks.
- Managed project lifecycles from initiation to deployment, contributing to **project roadmaps, planning,** and **requirement documentation** across cross-functional teams.
- Designed and maintained **data ingestion and transformation pipelines**, optimizing storage and compute for **Azure SQL, Synapse Analytics**, and **NoSQL** data stores.
- Monitored data pipelines and ETL systems for performance degradation, implemented alerting mechanisms, and resolved anomalies to ensure robust data delivery.
- Developed and enforced **data security protocols**, including access control, data encryption, and compliance with healthcare data protection regulations (e.g., HIPAA-aligned practices).
- Partnered with data scientists, analysts, and business stakeholders to translate analytical needs into **efficient and reliable data architecture**.
- Strong knowledge of **Relational Databases** with advanced **SQL** skills to proficiently write complex SQL queries for data analysis, development, and validation. good exposure of **UML**.
- Expertise in Data extraction, transformation, and loading (**ETL**) by creating dynamic packages using containers and various transforms in **SSIS**.
- Experience in using various packages in **Python** like NumPy, Pandas, Seaborn, Matplotlib, Beautiful Soup, etc, for **EDA**. And used **Pandas API** to put the data in a time series and tabular format for easy timestamp data manipulation and retrieval.
- Designed interactive visualizations using **Tableau**, **Power BI** software, and published and presented dashboards, storylines on web and desktop platforms.
- Good experience in **AWS** services like S3, IAM, EC2, EMR, Glue, QuickSight, SNS, SQS, Lambda, etc.
- Accomplished various tasks in a big data environment, which involved MS **Azure** Data Factory, Data Lake, and SQL Server.
- Involved in Daily standups and sprint planning, and review meetings in the **Agile** model.
- Extensively used **ERWIN** and Power Designer for data modeling techniques, star and snowflake schemas.

- Created a fully automated build and deployment platform and coordinated code builds, promotions, and orchestrated deployments using **GIT**.
- Used **JIRA** to build an environment for development.
- Worked on all stages of Software Development Life Cycle (**SDLC**).

ACADEMIC QUALIFICATIONS

Master of Science – MS, Business Analytics, Denton, Texas
University of North Texas – May 2025
Bachelor of Science in Electronics and Communication Engineering, Chennai, India
Anna University – May 2022

TECHNICAL SKILLS

Programming Languages	Python, Java, Scala, SQL, R, PostgreSQL
Cloud Platforms	AWS (S3, Glue, EC2, Lambda, Redshift), Azure (Data Factory, Synapse, Data Lake), GCP (BigQuery, Cloud Storage)
ETL & Data Pipelines	SSIS, Apache NiFi, AWS Glue, Azure Data Factory, Airflow
Data Warehousing	Snowflake, Amazon Redshift, Google BigQuery, Azure Synapse Analytics
Big Data Ecosystem Data Governance & Security	Apache Hadoop, Spark, Hive, Kafka IAM, data encryption, GDPR, HIPAA, RBAC for VDE, data use agreements, and IRB
Databases	PostgreSQL, MySQL, MS SQL Server, NoSQL (MongoDB, Cassandra)
Data Modeling	ERwin, PowerDesigner, Star/Snowflake Schema, UML
Big Data Ecosystem	Apache Hadoop, Spark, Hive, Kafka
Monitoring & Logging	CloudWatch, Grafana, Prometheus (basic), Datadog
Utilities & IDEs	Visual Studio Code, PyCharm, Jupyter Notebook, DBeaver, SQL Server Management Studio

PROFESSIONAL EXPERIENCE

Client: University of North Texas **September 2023 – Present**

Role: Data Engineer

Location: Texas

SEER-Medicare-Health Data Integration and Visualization Project

- Orchestrated the end-to-end integration of SEER cancer registry data, Medicare claims, and state-level public health datasets using ZIP and FIPS codes to unify disparate sources into a cohesive patient trajectory model from diagnosis through outcomes.
- Cleaned and preprocessed the data using SQL, PostgreSQL, and Python, standardizing formats and resolving inconsistencies across datasets.
  Designed and implemented robust **ETL pipelines** using **Apache NiFi**, automating ingestion and transformation processes.
- The cleaned data was stored in **Google Cloud Storage** and then loaded into **Google BigQuery**, which served as the central data warehouse. Built interactive dashboards and visual reports **using Metabase**, enabling stakeholders to explore trends, disparities, and outcomes across demographic and geographic dimensions. This solution facilitated evidence-based decision-making for healthcare researchers and policy planners.
- Unified Health Data Sources: Orchestrated the integration of SEER cancer registry, Medicare claims, and state-level health datasets using Zip and FIP codes, providing a comprehensive view of patient trajectories from diagnosis to outcomes.
- Enhanced Data Integrity: Addressed data quality issues by correcting missing values, eliminating duplicates, and standardizing column names to ensure a reliable dataset for analyzing breast cancer stages and survival times.
- Data Visualization Mastery: Utilized Matplotlib to produce scatter plots, bar charts, and heatmaps, revealing patterns and trends in the data to facilitate deeper insights into breast cancer research.
- In-depth Data Analysis: Employed SQL, Tableau, Excel, and machine learning tools to interpret visualized results, providing critical insights for research, reporting, and strategic oncology decision-making.
- Machine Learning Implementation: Applied various machine learning algorithms, such as Gradient Boosting Classifier and Ordinary Least Squares regression, to assess and correct biases in large-scale datasets, promoting model fairness.
- Predictive Model Development: Created predictive models using advanced machine learning methodologies to forecast cancer outcomes and responses to treatment.
- Transparency through Explainable AI/ML: Implemented explainable AI/ML approaches to ensure clarity and transparency in model predictions, aiding clinicians and researchers in decision-making.
- Rigorous Data Validation: Conducted thorough quality assurance and data validation to maintain the accuracy and reliability of the data utilized in analyses and modeling.
- Advanced Data Collection Techniques: Executed sophisticated data scraping methods to acquire extensive datasets from SEER, Medicaid, AHD, USDA, and CMS, vital for informed decision-making in health research.
- Contributions to Scientific Research: Played a key role in drafting research papers and communicating complex data analyses to the scientific community, enhancing understanding of public health impacts.
- Leadership in Cross-functional Collaboration: Worked closely with cross-functional teams to align data initiatives with broader research objectives, ensuring successful project execution and impactful results.

- Expanded Scope: Data Governance, Epidemiology, and Public Health Integration In addition to technical accomplishments, this project had a strong emphasis on data governance, epidemiological analysis, and public health impact.
- Data Governance Excellence: Developed and enforced policies and procedures for managing data privacy, access control, and compliance with HIPAA and other regulatory frameworks. Led the creation of a public health data governance framework that defined clear roles, responsibilities, and processes for ensuring data accuracy, consistency, and security.
- Epidemiological Analysis: Collaborated closely with epidemiologists to define cohort selection criteria, perform population-level health trend analysis, and calculate survival rates, disease incidence, and treatment response across various demographic and geographic groups.
- Public Health Impact: The project supported strategic decision-making for public health interventions by linking health data with geospatial and socioeconomic indicators. These insights were used to identify disparities in healthcare access and outcomes, guiding policy recommendations.

- Stakeholder Collaboration: Acted as a liaison between technical teams and public health officials, translating data requirements into scalable solutions and ensuring that analytical outputs aligned with real-world healthcare priorities.

- Transparency and Reproducibility: Maintained comprehensive documentation of data transformation processes, pipelines, and modeling decisions to support reproducibility and knowledge transfer.

Additional Contributions and Innovations

- Proactive Problem-Solving: Applied expert problem-solving skills to address and resolve technical and data-related issues as they arose, ensuring minimal downtime and maintaining high standards of data quality and system performance.
- Stakeholder Engagement and Communication: Regularly engaged with stakeholders through meetings and presentations, providing updates on project status, explaining complex technical details, and gathering feedback to refine data solutions. Developed comprehensive documentation to assist users and technical teams in understanding and utilizing the data systems.
- Innovative Technology Integration: Explored and integrated new technologies and methodologies to enhance the capabilities of the data infrastructure. This includes cloud computing solutions for scalable storage and processing, as well as advanced analytical techniques for predictive modeling and machine learning.
- Ethical Data Governance: Ensured that all data handling practices comply with ethical standards and legal requirements, particularly concerning data security and patient confidentiality. Implemented robust data governance frameworks to oversee the proper management of sensitive health data.

Social Service Data Project – Volunteer Data Specialist **January 2025 – Present**
Big Brothers Big Sisters | Code for Good
Remote January 2025 – present

- Partnered with nonprofit leadership to design a data-driven approach for tracking mentorship, fundraising, and volunteer outcomes.

- Engineered and managed relational databases using SQL, enabling streamlined access to program metrics and historical records.
- Developed automated ETL processes using SQL and Excel macros to extract, clean, and transform incoming data from diverse sources.
- Analyzed program trends using Python (Pandas, Seaborn, NumPy) and identified engagement patterns that led to improved outreach strategies.
- Built interactive dashboards and storyboards in Power BI and Tableau to visualize impact metrics for board meetings and donor reports.
- Collaborated with cross-functional teams (fundraising, program delivery, operations) to ensure data aligned with organizational goals.
- Created data quality assurance checks, reducing manual errors and improving reporting accuracy by 30%.
- Contributed to GitHub repositories for team code collaboration, version tracking, and documentation of SQL scripts and data models.
- Drafted internal guides and training materials to help nonprofit staff interpret dashboards and maintain basic queries.
- Supported quarterly performance reviews by delivering concise visual summaries and KPI analyses that influenced future programming.
- Participated in regular planning and Agile-style check-ins with NGO tech volunteers to ensure alignment of goals and timelines.
- Championed data ethics and privacy practices to ensure responsible handling of sensitive information in line with nonprofit standards.

Company: HCLtech **September 2022 - July 2023**
Role: Data Specialist/SQL Developer

Client: Toyota Financial Services

Worked as an IT Outsourcer

Location: Chennai, India

Responsibilities:

- **Strategic Data Management**: Orchestrated the automation of data extraction, transformation, and loading processes using **SSIS**, enhancing operational efficiency across Toyota’s global data systems.
- **Robust Data Integration and Quality Assurance**: Leveraged **Azure Data Factory** and **Python** scripting to cleanse and integrate large-scale data sets from diverse sources, ensuring high data quality and reliability for critical decision-making processes.
- **Advanced Analytics and Reporting**: Utilized **Power BI** and **SQL** to develop and administer insightful analytical reports and dashboards, directly supporting Toyota's strategic objectives by providing actionable insights into operational data.
- **Cloud Solutions Engineering**: Managed and optimized **Azure** cloud services, including **BLOB** and **Data Lake** storage, ensuring scalable and secure data solutions that support Toyota's outsourcing and IT strategies.
- **Automated Monitoring and Response Systems**: Implemented **Azure Automation** to monitor IT resources and configured alarms for proactive issue resolution, safeguarding system stability and performance.

- **Security and Compliance Protocols**: Conducted rigorous security assessments and compliance checks, aligning with Toyota's stringent data security requirements to protect sensitive information and system integrity.
- **Process Optimization and Agile Project Management**: Directed Agile project teams through daily scrums and sprint planning, driving the timely execution of IT projects that enhance Toyota's operational efficiencies.
- **Visualization and Decision Support Tools**: Crafted advanced visualizations using **Power BI**, **SSRS**, and **Excel**, facilitating strategic decisions with real-time data insights tailored to specific business needs.
- **Innovative Problem Solving and System Support**: Engaged in troubleshooting and resolving complex IT issues, improving system functionality and user satisfaction while providing continuous post-production support.
- **Agile Development & Collaboration:** Led cross-functional Agile teams, conducted daily scrums, sprint planning, and retrospective meetings to drive iterative delivery of data engineering tasks and ensure alignment with Toyota's IT roadmap.
- **Analytics Enablement & Visualization:** Built data marts and served curated datasets to BI teams; collaborated on **Power BI**, **SSRS**, and **Excel-based** dashboards to provide timely, actionable insights into supply chain, production, and financial KPIs.
- **Post-Production Support & Optimization:** Troubleshot complex data issues across environments, performed root cause analysis, and implemented automated data validation to reduce error rates and improve pipeline stability.
- **Infrastructure as Code & Deployment:** Utilized **ARM templates** for deployment of ADF pipelines, linked services, and datasets across dev, test, and prod environments, supporting scalable and repeatable infrastructure provisioning.

Environment & Tools: Azure Data Factory (ADF), Synapse Analytics, Azure Data Lake Storage, Blob Storage, SSIS, Power BI, SQL Server, T-SQL, Python, Azure Monitor, Azure Key Vault, ARM Templates, JIRA, SSRS, Excel, Git

---

Client: Edgerock **January 2020 – August 2022**
Role: Data Engineer
Location: Hyderabad, India (Remote)
Responsibilities:

- Cleaned, transformed, and loaded large datasets into cloud-based environments for analytics and reporting using Azure Data Factory and SQL Server
- Created automated validation workflows using Azure Functions and Python to clean CSV data, convert XML to JSON, and optimize file storage using Parquet format
- Built ETL pipelines using SSIS to process bulk Excel data, apply business rules, and log records into SQL Server for traceability and compliance
- Designed logging frameworks to track task-level and package-level ETL activities, ensuring high data quality and visibility
- Pre-aggregated data using Power BI Dataflows and Power Query to streamline dashboard generation and reduce report latency
- Utilized **MS Excel** (including **VLOOKUP**, pivot tables, and conditional formatting) and **MS Access** for ad-hoc analysis, error tracking, and offline reporting workflows

- Developed dashboards and visual reports in **Power BI** and **Tableau** to communicate KPIs and performance trends to stakeholders
- Migrated data integration workflows from Azure Data Factory V1 to V2, managing environment promotion using ARM templates and Azure Key Vault
- Managed **Azure** services such as Synapse, Blob Storage, and Data Lake for scalable data storage and transformation
- Wrote and optimized complex **T-SQL** queries, stored procedures, and views to support reporting, analytics, and automation
- Implemented row-level security and enforced data governance standards for access control across reporting layers
- Automated routine data ingestion tasks using shell scripts and integrated them into production schedules
- Partnered with QA and UAT teams to validate pipeline outputs and troubleshoot data issues in production
- Collaborated in Agile sprints and daily stand-ups to prioritize reporting deliverables and improve data processes

Key Tools & Technologies:
MS Excel, MS Access, Power BI, Tableau, SSIS, SQL Server, T-SQL, Azure Data Factory, Azure Synapse, Azure Blob Storage, Azure Data Lake, Python, Azure Monitor, ARM Templates, Git, JIRA, SSRS.

Company: One Eye Technology June 2019 - June 2019
Role: Intern
Location: Chennai, India.

- Assisted in collecting, cleaning, and transforming raw datasets from multiple sources using SQL, Excel, and Python, supporting data-driven decision-making processes.
- Developed basic ETL workflows under supervision, contributing to automated data pipelines using tools like SSIS or Azure Data Factory.
- Performed exploratory data analysis (EDA) and created summary reports using Power BI and Excel dashboards, highlighting key trends and metrics.
- Wrote and optimized SQL queries for data extraction, reporting, and performance testing in staging and production databases
- Collaborated with senior data professionals to understand data requirements, ensuring accurate data modeling and validation.
- Documented data definitions, data flows, and transformation logic to support transparency and knowledge transfer within the team.
- Gained hands-on experience in cloud platforms (Azure/AWS), including basic use of Data Lake Storage, Blob Storage, and Synapse Analytics.
- Participated in Agile ceremonies (stand-ups, sprint planning) and learned how to work within a structured software development lifecycle (SDLC).

CERTIFICATIONS:

Cloud: Azure Data Engineer Associate from
Microsoft, IoT (Internet of Things) Wireless & Cloud Computing Emerging Technologies

SQL: Analyze Data with SQL, From Excel to SQL, SQL Certifications, SQL for Marketers and Product Managers, Design Databases with PostgreSQL

Artificial Intelligence: Copilot AI, Generative AI & Azure Bot AI Service, Data and Programming for AI
BI: Business Intelligence Data Analyst, Data Scientist Analytics, Data Scientist Natural Language Processing Specialist, Data Engineer, Data Scientist, Inference Specialist, Data Science Foundations, Power BI
Programming: Python for Data Science, Crash Course on Python, Java for Programmers, Intermediate R, Introduction to Programming with MATLAB, **Information Security**: Cybersecurity

---

Real Time project:

1. **Health Data Integration and Management System**

- Designed and deployed SQL-based databases on Microsoft Azure to manage and store large health data, ensuring secure, efficient access across departments.
- Developed ETL pipelines using Python and SQL, automating data extraction, transformation, and loading processes for improved data accuracy and consistency.
- Ensured HIPAA compliance by implementing secure data access controls, encryption protocols, and safe data sharing practices across health systems.
- Created automated data quality checks to monitor and maintain high data integrity, identifying and resolving discrepancies in incoming datasets.
- Collaborated with data scientists and IT teams to integrate data from multiple sources and develop reporting tools, streamlining decision-making.
- Built interactive Power BI dashboards to visualize health trends, patient outcomes, and key metrics, enabling real-time insights for stakeholders.
- Reduced manual data entry by 40% and improved reporting accuracy by automating data processing workflows.
- Improved data processing efficiency and ensured data integrity through the implementation of automated ETL processes.

2. **Patient Health Records ETL Pipeline**

- Goal: Built an end-to-end pipeline to ingest, clean, and store electronic health records (EHR) data.
- Tech Stack: Azure Data Factory / SSIS, SQL, Python, Azure Data Lake, Power BI
- Ingest EHR data (CSV/XML/JSON) from clinics or hospitals. Use Python to clean and standardize diagnosis codes, dates, and missing values. Store raw and processed data in Azure Data Lake. Load into a star-schema model in Azure Synapse Analytics. Build a Power BI dashboard showing patient trends (e.g., readmissions, common diagnoses).
- Integrated geospatial data into analytics pipelines using GeoPandas and ArcGIS, enabling spatial insights for public health surveillance and resource allocation.

3. **Real-Time Health Monitoring Data Pipeline**

- **Goal:** Streamed and processed real-time wearable health data (heart rate, SpO2, etc.).
- **Tech Stack:** Azure Event Hubs / Kafka, Azure Stream Analytics / Spark Streaming, Azure SQL / Synapse

- Simulate or collect data from wearable devices. Stream the data to Event Hubs or Kafka. Use Stream Analytics or Spark to clean and aggregate the data. Store processed data in Azure SQL or Synapse. Trigger alerts if vitals cross risk thresholds.

4. Electric Vehicle (EV) Data Analysis and Forecasting

- Performed exploratory data analysis (EDA) to identify key trends and patterns in EV adoption and usage data.
- Built and optimized Random Forest and XGBoost models to predict critical factors impacting EV growth and performance.
- Developed a time series forecasting model using Prophet to forecast future EV adoption trends and market demand.
- Conducted feature engineering and hyperparameter tuning to enhance model accuracy and robustness.
- Visualized insights through interactive plots and charts to support data-driven decision-making.
- Tools: Python, Pandas, Scikit-learn, XGBoost, Prophet, Matplotlib

5. Healthcare Patient Data Analysis and Forecasting

- Analyzed healthcare datasets containing patient demographics, diagnosis codes, treatments, and outcomes.
- Performed exploratory data analysis (EDA) to identify trends in hospital admissions, disease prevalence, and patient readmissions.
- Built Random Forest and XGBoost models to predict patient readmission risk and disease occurrence based on clinical variables.
- Applied Prophet time series forecasting to predict future hospital admission rates and patient volume trends.
- Visualized key insights through charts and dashboards to assist in healthcare resource planning and decision-making.
- Technologies: Python, Pandas, Scikit-learn, XGBoost, Prophet, Matplotlib, Seaborn

---

Case Study: Financial Impact Analysis of Drone Delivery Technology for Kroger

- Tools & Technologies: Python (Pandas), Excel, Financial Modeling, Drone Technology, AI, Geospatial Data, Data Analysis
- Conducted a comprehensive financial analysis for Kroger to assess the potential revenue growth and cost savings from implementing drone-based last-mile delivery for small grocery orders.
- Analyzed Kroger's fiscal data and projected a 5% increase in e-commerce revenue with drone adoption, estimating an incremental revenue of $500 million over the next 5 years.
- Developed a financial model to forecast drone delivery operational costs, revenue growth, and cost savings in delivery fleet operations (projected 20% reduction in fleet costs).
- Evaluated the regulatory, public perception, and operational challenges in adopting drone technology, ensuring compliance with airspace regulations.

- Presented findings to senior management, highlighting the strategic advantages of adopting drone delivery to enhance customer satisfaction, increase revenue, and optimize operational costs.