Venkata Gupta Penugonda

Data Engineer

☎ 206-424-6482 ✉ venkataguptapenugonda@gmail.com

Professional Summary

Experienced Data Engineer with over 5 years of expertise in designing and implementing scalable data solutions across Azure, AWS, and GCP cloud platforms. Proven track record of developing robust ETL/ELT pipelines, real-time streaming architectures, and cloud-native data lakehouses for clients in finance, insurance, healthcare, and pharmaceutical domains. Skilled in tools such as Azure Data Factory, AWS Glue, Apache Airflow, Databricks, Spark, and BigQuery. Strong knowledge of data governance, quality frameworks, and compliance with HIPAA and SOX standards. Hands-on experience with DevOps practices including CI/CD automation using Azure DevOps, Jenkins, Terraform, and CloudFormation. Adept at collaborating with cross-functional teams in Agile environments to deliver high-performance, secure, and business-aligned data products.

Education

University of South Florida **Florida, USA**
Masters in Computer Science

Experience

Bank of America **Oct 2023 - Present**
Senior Data Engineer *Cleveland, Ohio*

- **Enterprise Cloud Modernization for Banking Data Lake** – Migrated legacy batch workflows to a cloud-native data lakehouse architecture on **Azure**, centralizing enterprise financial and transactional data.
- Led the end-to-end development of scalable **Azure Data Factory (ADF)** pipelines, orchestrating ingestion of structured and semi-structured data from over 15 banking systems into **Azure Data Lake Storage (ADLS)** and **Azure SQL Database**.
- Designed and implemented a modular **Delta Lake** architecture using **Azure Databricks**, enabling ACID transactions, schema enforcement, and optimized time-travel query support for historical financial data.
- Improved batch ETL performance by 45% by refactoring legacy SSIS packages into **ADF pipelines** with dynamic parameterization and integration with **Azure Key Vault**.
- Developed **PySpark** scripts in **Azure Databricks** for cleansing, transformation, and aggregation of high-volume transactional data supporting risk and fraud analytics teams.
- Enabled near real-time processing of streaming data using **Azure Event Hubs**, **Azure Stream Analytics**, and **Synapse Analytics**, reducing fraud detection latency from hours to under 5 minutes.
- Built robust CI/CD pipelines for data workflows using **Azure DevOps**, automating deployments across dev, UAT, and prod environments with environment-specific ARM templates.
- Collaborated with data governance teams to implement data lineage, cataloging, and classification in **Microsoft Purview**, ensuring compliance with internal controls and **SOX** requirements.
- Developed and optimized complex **T-SQL** and **Spark SQL** queries to support reporting and regulatory submissions, reducing overall execution time by 30%.
- Actively participated in Agile sprints, working closely with cross-functional teams including compliance, cybersecurity, and data science to deliver resilient and scalable data products aligned with KeyBank's cloud modernization goals.

Root Insurance **Nov 2022 - Sep 2023**
AWS Data Engineer *Tampa, Florida*

- **Real-Time Telematics and Claims Data Platform** – Engineered an end-to-end AWS-based solution to process vehicle telemetry and claims data for fraud detection, underwriting, and personalized policy pricing.
- Designed and implemented scalable **ETL pipelines** using **AWS Glue**, ingesting and transforming JSON/CSV/parquet data from S3, API Gateway, and third-party telematics platforms.
- Built near real-time ingestion and processing flows using **Kinesis Data Streams**, **Lambda**, and **AWS Glue Streaming**, reducing claim fraud detection latency from hours to under 10 minutes.
- Optimized **Athena** and **Redshift Spectrum** queries for analytical workloads, improving performance by 35% for data scientists and actuaries consuming vehicle and behavioral risk models.
- Implemented robust job orchestration using **Apache Airflow on MWAA**, managing over 80 DAGs across dev, QA, and prod environments for data quality, batch updates, and machine learning feature generation.
- Built infrastructure-as-code using **Terraform** and **CloudFormation**, ensuring consistent provisioning of data pipelines, IAM roles, S3 buckets, and KMS-encrypted resources.

- Established data governance and access control using **Lake Formation**, **Glue Catalog**, and tag-based policies to maintain HIPAA and PCI compliance.
- Monitored and debugged real-time and batch pipelines using **CloudWatch**, **X-Ray**, and **CloudTrail**, reducing SLA breaches by 25% and improving observability.
- Collaborated across data science, product, and actuarial teams to design high-availability architecture supporting ML-based risk scoring for 2M+ customers.
- Worked in Agile sprints, contributed to sprint planning, retrospectives, and feature refinement for continuous integration and incremental delivery of data platform capabilities.

Novo Nordisk **Jun 2021 - Jul 2022**
Data Engineer *Hyderabad, India*

- **Clinical Trial Analytics and Pharma Data Integration** – Developed a hybrid data engineering solution supporting global trial data processing, patient adherence tracking, and regulatory submissions.
- Built automated batch pipelines using **Apache Airflow** and **Python**, integrating EHR and clinical trial data from flat files, APIs, and relational databases into a centralized data warehouse.
- Leveraged **Google Cloud Storage (GCS)**, **Cloud Composer**, and **BigQuery** for secure storage and querying of anonymized patient datasets to support pharmacovigilance analytics.
- Developed ETL jobs using **Informatica** and custom **PySpark** scripts for transforming trial protocol, adverse event, and medication data into standardized formats (CDISC, SDTM).
- Enabled self-service reporting and dashboarding for clinical stakeholders by integrating curated datasets into **Looker** and **Power BI**, improving data accessibility and decision-making.
- Worked closely with GCP security teams to enforce access policies using **IAM roles**, service accounts, and VPC-SC configurations for compliant healthcare data handling.
- Deployed ML-ready datasets to **BigQuery ML** for running patient dropout prediction models in collaboration with the data science team.
- Implemented **Data Quality Frameworks** for consistency checks across trial phases using dynamic rule engines and validation rules stored in metadata tables.
- Conducted daily Agile ceremonies and worked alongside clinical data managers, statisticians, and GCP architects to ensure timely delivery of analytics-ready datasets.

Bayer **Mar 2020 - May 2021**
Junior Data Engineer *Hyderabad, India*

- **Crop Analytics & Pharmaceutical Supply Chain Data Platform** – Supported the development and maintenance of batch ETL processes to unify data from research labs, field trials, and production systems.
- Assisted in building ETL pipelines using **Talend** and **Python**, enabling the extraction and integration of crop genetics, chemical trial data, and logistics datasets from multiple silos into enterprise warehouses.
- Supported ingestion of clinical product data from SAP and LIMS systems into **SQL Server** and **PostgreSQL** databases using data quality and mapping rules for consistency.
- Developed and maintained scheduled jobs via **Apache NiFi** and **Talend JobServer** to automate file transfers and validation processes across research locations.
- Created basic **data validation scripts** in **Python** to perform QA checks on raw datasets for trial duration, batch formulation, and shipment tracking data.
- Worked closely with domain SMEs and senior engineers to support the migration of batch workflows into a centralized **AWS S3**-based archive for historical analysis.
- Built internal dashboards using **Power BI** to provide visibility into trial completion timelines, active SKUs, and regional shipment volumes.
- Documented pipeline logic, metadata flows, and SOPs in **Confluence** to support reproducibility and audit readiness for compliance purposes.
- Participated in daily standups and weekly sprint reviews under the guidance of tech leads and product owners to deliver backlog items and enhancements.

Technical Skills

Languages: Python, SQL, Java, Shell Scripting, HTML/CSS, JavaScript
Cloud Platforms: Microsoft Azure (ADF, ADLS, Synapse, Databricks, Event Hubs, Purview), AWS (Glue, Redshift, Lambda, Kinesis, S3, CloudWatch, IAM), GCP (BigQuery, GCS, Cloud Composer, BigQuery ML)
ETL/Orchestration Tools: Azure Data Factory, AWS Glue, Apache Airflow, Talend, Informatica, Apache NiFi
Big Data & Processing: Apache Spark, PySpark, Databricks, Delta Lake, Hadoop
DevOps & CI/CD: Azure DevOps, GitHub, Jenkins, Terraform, CloudFormation, Docker
Data Warehousing: Azure Synapse Analytics, Amazon Redshift, Google BigQuery, SQL Server, PostgreSQL
Data Visualization: Power BI, Looker, Tableau
Data Governance: Microsoft Purview, AWS Lake Formation, Glue Catalog, IAM, HIPAA, SOX
Developer Tools: Visual Studio Code, Eclipse, Jupyter Notebook, Postman, Confluence
Operating Systems: Linux, Windows