REVANTH GOSULA
Data Engineer
(513) 257-8371 | revanthgosula5@gmail.com **|** LinkedIn **| Cincinnati, OH**

PROFESSIONAL SUMMARY

Experienced **Data Engineer** with **3+** years of experience in developing enterprise-grade **data pipelines**, real-time workflows, and **data lake architectures** across financial, healthcare, and consulting domains. Proficient in **Python, SQL, Spark, Kafka, dbt, Talend, Snowflake, PostgreSQL, MySQL, Oracle,** and **Informatica**. Skilled in cloud platforms such as **Azure, AWS**, and **Snowflake**. Expertise in Agile environments using **JIRA** and **Git**, with strong knowledge of **data governance**, security, and compliance. Generated model-ready datasets and supported **GenAI** use cases through vector database prep, prompt-tuning, and retrieval-augmented generation (**RAG**) pipelines with **LLMs**. Experienced in **MLflow**, **REST APIs**, **Terraform**, **Jenkins**, **Git**, **Power BI**, and **Tableau**. Aiming to apply technical expertise to drive transformative data solutions and serve business objectives in a forward-thinking organization.

TECHNICAL SKILLS

- **Data Analysis:** SQL, TSQL, Python (Pandas, NumPy, SciPy), R, Excel
- **Big Data Technologies:** Apache Spark, Hadoop, Hive, HDFS, Kafka, Databricks
- **ETL & Data Integration:** Apache NiFi, Talend, AWS Glue, dbt, Azure Data Factory, Informatica, Fivetran, Airbyte
- **Data Warehousing:** Snowflake, Amazon Redshift, Azure Synapse, Google BigQuery
- **Cloud Platforms:** AWS (S3, Glue, EMR), Azure (Blob, Synapse, ML Studio), Google Cloud, GCS, Vertex AI
- **Databases:** Snowflake, MySQL, PostgreSQL, Oracle, Microsoft SQL Server
- **Data Visualization:** Power BI, Tableau, Looker
- **Machine Learning & AI:** scikit-learn, TensorFlow, RAG, Hugging Face Transformers, OpenAI API, LangChain, Prompt Engineering, Feature Engineering, Vector Databases, Embedding Pipelines, MLflow
- **DevOps & Automation:** Airflow, Jenkins, Docker, Kubernetes, Terraform, Git
- **Version Control & CI/CD:** Git, GitHub, Bitbucket
- **Data Security & Compliance:** HIPAA, GDPR, Data Masking, Encryption, IAM
- **Other Tools:** REST APIs, JSON, XML, JIRA, Agile/Scrum, SDLC

PROFESSIONAL WORK EXPERIENCE

Data Engineer **Dec 2024 – Present**
Citi | Remote

- Architected **real-time data ingestion pipelines** with **Kafka**, **Spark Structured Streaming**, and **Airflow**, processing millions of events per hour across Citi's equities, fixed income & FX desks.
- Engineered data integration pipelines with **Talend & Informatica** to load financial data into **Snowflake**, elevating audit compliance and report accuracy by **30%.**
- Developed an **AWS data lake solution** with **S3, Glue, and Athena,** managing **Parquet/ORC** datasets with **JSON/XML** schema versioning to assist **AI/ML** workflows including **LLM** fine-tuning.
- Delivered **ML-powered** dashboards in **Power BI** and **Tableau,** leveraging **scikit-learn** and tuned Snowflake queries, accelerating insight speed and increasing user adoption by **65%.**
- Streamlined cross-source data processing with **Python**, **SQL**, and **Spark**, consolidating structured and unstructured datasets from **Oracle, MySQL**, and **PostgreSQL**.
- Implemented scalable data workflows in **AWS (S3, Glue, Lambda)** with **Airflow** for orchestration and **terraform** for infrastructure provisioning, decreasing **ETL** runtime by **40%.**
- Created complex **T-SQL** procedures supporting regulatory reporting and ML model training data preparation, enhancing **data accuracy** and compliance with regulatory standards.
- Refined **Snowflake** data modeling with **clustering** and **caching**, enriching query performance by **30%** for **BI** and **ML** workflows, and configured feature validation with **Hugging Face** Transformers.
- Created modular, testable data models in **Snowflake** using **dbt**, mitigating model debugging time by **40%** and accelerating regulatory report delivery across **5+** financial products.
- Streamlined **data governance**, quality, and compliance by implementing profiling, validation, and metadata management across **Hadoop, Kafka,** and **API** data sources, increasing data accuracy and reliability.
- Led a **cross-functional** team of data analysts and engineers during a sprint to troubleshoot and optimize a trade data pipeline, reducing processing delays and amplifying **SLA** adherence by **25%.**

Environment: Apache Kafka, Apache Spark, Apache Airflow, Talend, Informatica, Snowflake, AWS (S3, Glue, Athena, Lambda), PostgreSQL, Python (Pandas, NumPy, TensorFlow), SQL, Power BI, Tableau, scikit-learn, MLflow, Docker, Jenkins, Hugging Face Transformers, OpenAI API

Data Engineer **Nov 2022 – Nov 2023**
Deloitte | India

- Engineered batch and real-time **ETL** pipelines with **Apache Spark** and **Kafka** for **Basel III** and **IFRS-9** regulatory data, creating **ML**-ready datasets for anomaly detection and compliance scoring.
- Implemented **RAG**-based document search systems using **vector databases** and **databricks** and embedding models for regulatory document retrieval and compliance automation.
- Constructed scalable **ETL** flows using **Informatica** and **Python (scikit-learn)**, transforming raw credit and loan data into validated schemas for downstream analytics.
- Migrated legacy workloads to **Azure Data Factory**, boosting pipeline load times by **30%** and integrating **MLflow** for pipeline performance and data quality monitoring.
- Programmed **data quality** validation with **Python** and **Excel macros**, cutting QA time by **50%** and helping anomaly detection using **TensorFlow**-based outlier models.
- Enhanced large-scale ML preprocessing efficiency using **Databricks** notebooks and synced outputs to **Azure Synapse**, resulting in strengthened unified reporting.
- Designed **Tableau** dashboards to visualize credit risk, portfolio exposure, and capital metrics by integrating curated datasets from **Oracle** and **PostgreSQL.**
- Streamlined scalable **data** pipelines using **Python**, **SQL**, and **Apache Spark** across **Oracle**, **MySQL**, and **PostgreSQL**, automating feature scaling and encoding for **ML** consumption.
- Built scalable **ETL** workflows in **Azure Data Factory (ADF)** and orchestrated pipelines using **Azure DevOps** and **Git**, aligning with **DevOps** standards for version control and continuous integration to fasten deployment time by **25%.**
- Utilized **Informatica** and **Talend** for data extraction, cleansing, and transformation across multiple business domains, ensuring data integrity and consistency across layers.
- Delivered efficient **MLOps** pipelines using **GCP Vertex AI, MLflow**, and **Kubeflow**, minimizing model training, validation, and deployment cycles from weeks to mere hours.
- **Collaborated** with risk analysts and data engineers to fine-tune **IFRS-9** pipelines, improving **data accuracy** for regulatory reporting and helping team meet audit deadlines.

Environment: Apache Spark, Kafka, Informatica, Python (Pandas, NumPy, scikit-learn, TensorFlow), Azure Data Factory, GCP, Azure DevOps, Snowflake, Oracle, PostgreSQL, MySQL, SQL Server, Tableau, MLflow, REST APIs, JSON/XML

Data Engineer **Jan 2021 – Oct 2022**
Optum | India

- Automated healthcare **ETL** jobs using **Azure Data Factory triggers** and **parameterized pipelines**, lowering manual intervention and boosting **SLA** adherence for monthly claims data refreshes.
- Aligned data ingestion and transformation workflows with **JIRA,** business needs in **Agile** sprints, and daily stand-ups.
- Designed **ADF pipelines** to process patient claims and clinical records, bringing down data prep time by **35%** and allowing advanced analytics and model-ready outputs.
- Leveraged **Azure Synapse Analytics** to process large-scale medical records, advancing data for downstream **Power BI** reporting on patient outcomes, diagnosis trends, and policy utilization.
- Collaborated with data analysts and care management teams to deliver curated datasets using **Azure Blob Storage, Azure SQL** and **REST APIs**, ensuring compliance with **HIPAA** regulations.
- Developed and upgraded **T-SQL** scripts and real-time **ETL** workflows across **SQL Server, PostgreSQL**, and **MySQL**, maximizing data processing speed by **35%** and aiding faster downstream analytics.
- Designed scalable **ETL** pipelines using **Python (NumPy, Pandas)** and **SQL** across PostgreSQL, MySQL, and SQL Server, refined for real-time insights and **machine learning** workflows with orchestrated feature validation.
- Orchestrated daily data workflows using **Azure Data Factory** and **Azure DevOps**, integrating **Git** for version control, **CI/CD** automation, and **MLflow** for experiment tracking across healthcare **ML pipelines.**
- Constructed **data marts** and optimized queries in **Snowflake** and **SQL Server**, empowering real-time clinical dashboards and feature stores for advancing **ML** inference in healthcare analytics.
- **Coordinated** priorities between data engineering and clinical reporting teams to ensure timely ingestion of patient records, enabling accurate and on-time delivery of outcome dashboards.

Environment: Azure Data Factory, Azure Synapse Analytics, Azure Blob Storage, Azure SQL, Python (Pandas, NumPy, scikit-learn, TensorFlow), SQL Server, PostgreSQL, MySQL, Power BI, Tableau, MLflow, Git, REST APIs, JSON, XML, Hugging Face Transformers

EDUCATION

- **Master of Science in Information Technology** from University of Cincinnati, Cincinnati, OH, USA.
- **Bachelor of Science in Information Technology** from SASTRA Deemed University, Thanjavur, India.