Revanth Sai

revanthsai276@gmail.com

☎ +1(602)-921-7089

Professional Summary

• Data Engineer with 5+ years of strong expertise in designing and maintaining **end-to-end ETL pipelines** using Python across cloud and on-prem environments.
• Proficient in building scalable, reliable, and high-performance data pipelines for both batch and streaming data workflows using tools like **Apache Airflow**, **Azure Data Factory**, and **GCP Dataflow**.
• Advanced knowledge of **Python programming**, including data manipulation with **Pandas**, **NumPy**, and integration with cloud services and APIs.
• Extensive experience in **data modeling, transformation, and ingestion** into modern cloud data warehouses like **BigQuery**, **Snowflake**, and **Legacy databases**
• Strong SQL and NoSQL skills, with experience in **MySQL, MongoDB, and Cosmos DB**, including query optimization and schema tuning.
• Well-versed in **data quality frameworks**, implementing validations, checks, and exception handling across pipeline stages.
• Skilled in debugging complex data issues, automating monitoring, and documenting workflows to ensure transparency and maintainability.
• Collaborative problem solver with experience partnering with analysts, data scientists, and business stakeholders to deliver data-driven solutions.
• Exposure to AI/ML use cases through collaboration on model-ready data pipelines and academic AI specialization.

Technical Summary

Languages: Python, SQL, Shell Scripting, PySpark
ETL/Orchestration: Apache Airflow, Azure Data Factory, GCP Dataflow, SSIS, Azure Synapse
Cloud Platforms: GCP (BigQuery, Pub/Sub, Composer), Azure (Data Lake, Synapse), **DataProc, Google Composer**
Data Warehouses: BigQuery, Snowflake, Teradata **and** Legacy Databases
Databases: MongoDB, Cosmos DB, Cassandra, MySQL, Oracle
Big Data Technologies: Databricks, Hadoop, Kafka, **DataProc**
Data Modeling: Star/Snowflake Schema, ERD, Normalization, Partitioning, Clustering
CI/CD & DevOps: Git, GitHub, Jenkins, Docker, Azure DevOps
Data Governance & Compliance: HIPAA, GDPR, Audit Documentation
Visualization Tools: Power BI, Tableau

Professional Experience

Data Engineer

Verizon — *Feb 2023 – Present*

Project Summary –

Verizon: GCP-Based Customer & Operations Analytics Platform
At Verizon, I led the migration of 12+ legacy Teradata ETL workflows to **GCP BigQuery**, improving query speed by **50%** and reducing monthly compute costs by **30%**. I built real-time ETL pipelines using **GCP Dataflow (Apache Beam)** to ingest over **2M daily events** from customer usage, billing systems, device logs, and call center transcripts.

Pipelines were orchestrated via **Airflow (Cloud Composer)** with error handling, retries, and monitoring, reducing failure rates by **40%**. I implemented Python-based data validation that cut manual QA by **60%**, and designed **partitioned BigQuery tables** for sub-second querying.

Key **Power BI dashboards** included:
- **Churn Prediction** – tracked 250K+ users weekly to identify high-risk accounts.
- **Network Health** – visualized signal drops and outage patterns across 150+ regions.
- **Support Insights** – analyzed sentiment from 1M+ support chat logs monthly.

I also enabled **CI/CD** with Docker and GitHub Actions, and delivered **model-ready datasets** used in ML-driven personalization and fraud detection pipelines, boosting model performance by **20%.**

- Developed and managed scalable **ETL pipelines using GCP Dataflow**, processing structured and unstructured data into BigQuery.
- Rewrote legacy ETL workflows during **Teradata to BigQuery migration**, optimizing SQL logic and enhancing performance by 30%.
- Implemented Airflow DAGs (Cloud Composer) for orchestrating batch and real-time pipelines with error handling and retry logic.
- Built modular Python scripts for **data validation**, schema enforcement, and logging, improving pipeline reliability.
- Created optimized BigQuery schemas with clustering and partitioning, enabling sub-second query performance.
- Applied SQL and Python-based unit tests to verify pipeline outputs and catch data anomalies early.
- Created SSIS packages and integrated legacy workflows with new cloud-native pipelines.
- Collaborated with visualization team for Power BI dashboards using curated datasets and built DAX measures for KPI tracking.
- Documented all ETL workflows, data dictionaries, and governance policies in Confluence.
- Supported CI/CD automation using GitHub Actions and Jenkins, containerizing jobs with Docker.
- Worked closely with analytics and ML teams to provide **model-ready datasets** for personalization and forecasting.

Environment: Python, Airflow, BigQuery, GCP Dataflow, SSIS, SQL, Docker, Git, Power BI, MongoDB

Data Engineer

Humana — *Aug 2021 – Sep 2022*

Project Summary – Humana: Azure-Based Healthcare Data Modernization

At Humana, I designed and deployed scalable ETL pipelines using **Azure Data Factory** and **Databricks**, transforming healthcare claims, eligibility, and clinical records for over **10M+ members**. These pipelines supported downstream analytics, risk modeling, and operational reporting across actuarial and care teams.

I implemented **PySpark**-based transformation logic and modular **Python scripts** for schema validation and anomaly detection, improving pipeline reliability by **35%**. I built **star-schema models** in **Azure Synapse**, enabling performant query execution and dashboarding across large datasets.

Key Power BI dashboards included:
• **Care Utilization Tracker** – visualized patient visits, treatment types, and frequency patterns across provider networks.
• **Cost & Claims Insights** – monitored per-member cost trends, fraud indicators, and claims cycle times.
• **Chronic Risk Scoring** – enabled proactive care via population-level risk segmentation for 5M+ chronic care patients.

I also ensured HIPAA compliance, enabled **CI/CD via Azure DevOps**, and partnered with data scientists to deliver **ML-ready datasets** that improved model precision by **25%**.
• Designed and deployed **data pipelines in Azure Data Factory** and **PySpark** to support healthcare claims processing and clinical analytics.
• Built modular Python scripts for transformation logic, metadata checks, and pipeline error reporting.
• Used Azure Synapse to design **star-schema models** and write optimized T-SQL for reporting layers.
• Established reusable pipeline templates with data quality checks embedded for ingestion from external vendors.
• Automated pipeline testing using Python and PyTest, improving QA efficiency.
• Collaborated with data scientists to **prepare AI/ML datasets**, ensuring data consistency and feature readiness.
• Tracked performance using Spark UI and Azure Monitor, reducing job latency by 25%.
• Managed secure data access using role-based permissions, aligning with HIPAA requirements.
• Used Azure DevOps for CI/CD workflows and code versioning across development environments.
• Documented pipelines, schemas, and governance controls for audit readiness.

Environment: Azure Data Factory, Synapse, PySpark, SQL, Python, Azure DevOps, Cosmos DB, Power BI

Data Engineer

Warner Music Group — *Jun 2019 – Aug 2021*

Project Summary – Warner Music Group: GCP-Based Streaming Analytics Platform

At Warner Music Group, I built real-time and batch data pipelines using **Databricks**, **Airflow**, and **BigQuery**, processing over **500M monthly streaming events** from Spotify, YouTube, and internal platforms. These pipelines supported artist analytics, royalty forecasting, and global revenue reporting.

I migrated legacy ETL from Oracle to **GCP**, leveraging **DataProc**, **Pub/Sub**, and **BigQuery**, reducing job latency by **40%** and enabling faster access to trend data. I developed **Python scripts** for ingesting partner metadata and cleaned unstructured event logs for downstream use.

Key Power BI and Tableau dashboards included:
• **Fan Engagement Trends** – tracked song plays, skips, and playlist adds across regions and platforms.
• **Global Revenue Insights** – visualized top-performing artists, albums, and territories in near real-time.
• **Campaign Performance Monitor** – evaluated marketing effectiveness across digital campaigns and social platforms.

I also enabled **ML datasets** used for forecasting hit potential, audience clustering, and fan retention modeling, increasing campaign ROI visibility by **30%**.
• Engineered batch and streaming pipelines using **DataProc**, Google Cloud Composer, and **GCP services** for sales and consumption analytics.
• Rebuilt ETL jobs from Oracle SQL into **GCP BigQuery + DataProc** stack, improving performance and scalability.
• Developed Python utilities for ingesting API-based and file-based sources into Snowflake and BigQuery.
• Built Airflow DAGs with Python Operators to automate daily and hourly refreshes.

Developed scalable container utilities for data processing and deployed on Google Kubernetes Engine.
• Enabled Do
• Worked with stakeholders to define **data models** that supported key business metrics and external reporting needs.
• Performed anomaly detection and pipeline health monitoring using Python scripts.
• Assisted in log pipeline redesign, migrating from Cassandra to Snowflake for improved analytical performance.
• Collaborated for executive-facing dashboards in Power BI and Tableau, sourced from processed pipeline outputs.
• Provided training on pipeline operations and documentation to analysts and junior engineers.

Environment: DataProc, Google Cloud Composer, BigQuery, Snowflake, Spark, Docker, Python, Tableau, Power BI

Education

Master of Science in Information Technology Management (AI Specialization)
University of Wisconsin-Milwaukee — *Dec 2023*