Vedhasree Bonala

Little Rock, AR
5019442776 vedasri996@gmail.com www.linkedin.com/in/vedhasreeb https://vedhabonala.github.io/

Professional Summary

Data Engineer with 4 years of experience building scalable, cloud-native data pipelines and analytical models across AWS, Azure, and GCP. Skilled in designing ETL/ELT workflows, optimizing SQL and PySpark transformations, and developing curated data layers that power BI dashboards and advanced analytics. Experienced in data lakes, Lakehouse architectures, streaming ingestion, and data quality frameworks. Proven ability to partner with BI, analytics, and product teams to deliver reliable, high-performing data solutions that drive business decisions.

Work Experience

UnitedHealth Group | Data Engineer (Contract) | Remote **Jan 2025 - Present**

- Engineered HIPAA-compliant AWS ELT pipelines using Glue, Lambda, and PySpark to process large clinical, claims, and eligibility datasets, improving pipeline reliability and throughput.
- Designed and maintained S3-based data lake layers with automated ingestion, schema evolution, and optimized partitioning tailored for healthcare analytics and actuarial workflows.
- Implemented Redshift transformation logic using distribution/sort keys, query pruning, and workload management (WLM), resulting in significantly faster execution of complex actuarial/claims queries.
- Built data quality frameworks covering validation rules, anomaly detection, referential integrity checks, and reconciliation routines, reducing downstream reporting issues across clinical reporting teams.
- Developed metadata-driven ingestion patterns with auditability, lineage tracking, and compliance controls for PHI/PII-sensitive datasets in alignment with HIPAA security standards.
- Optimized PySpark workloads for healthcare transformations using encryption, masking, IAM-based access controls, and secure storage patterns.
- Delivered curated, analytics-ready data marts powering actuarial models, provider insights, utilization management, and executive reporting dashboards.
- Partnered with clinical analytics, BI, and product teams to translate requirements such as claims adjudication, benefits processing, and enrollment trends into scalable, production-grade data solutions.

Arkansas Summer Research Institute | Data Engineer Intern | Little Rock, *Arkansas* **May 2024 - Nov 2024**

- Developed Azure-based ETL/ELT pipelines using ADF, Databricks, and PySpark to process multi-format research datasets, improving pipeline efficiency and reliability.
- Implemented Lakehouse architecture using ADLS and Delta Lake with curated bronze, silver, and gold layers to support analytics and machine learning workflows.
- Built automated ingestion pipelines in ADF integrating APIs, Blob Storage, and SQL sources with monitoring, error handling, and operational alerting.
- Modeled analytical layers in Synapse using star-schema and incremental load patterns to optimize query performance for research teams.
- Built interactive Tableau dashboards enabling research groups to explore experimental data and statistical trends with improved clarity.

Wipro | Data Engineer | Hyderabad, *India* **Aug 2022 - Dec 2023**

- Engineered large-scale ETL pipelines using PySpark, Databricks, and AWS Glue to ingest, transform, and curate multi-terabyte enterprise datasets, improving overall pipeline performance and reliability.
- Designed and optimized dimensional models in Snowflake and Redshift (star and snowflake schemas) to support analytical dashboards used by finance, operations, and BI teams.
- Developed scalable ingestion frameworks orchestrated through Apache Airflow to automate cross-source loads from APIs, RDS, S3, and streaming systems, reducing repetitive manual workflows.
- Improved PySpark transformations using partitioning, broadcast joins, caching, and compression techniques, resulting in smoother large-scale processing.
- Processed enterprise datasets using Hadoop, Hive, and HDFS to support big data pipelines used in compliance, reporting, and operational analytics.
- Executed Spark-based workloads on AWS EMR for heavy transformation tasks, enabling scalable and cost-effective processing across big data environments.
- Designed and fine-tuned SQL transformations and stored procedures across PostgreSQL, Snowflake, and Redshift to support analytics, KPI reporting, and operational pipelines.
- Implemented data quality validations such as null profiling, anomaly detection, and referential integrity checks within Airflow-managed workflows, ensuring consistent and accurate output for BI teams.

Verizon | Data Engineer | Chennai, *India* **Nov 2021 - Aug 2022**

- Developed GCP-native ETL pipelines using BigQuery, Dataflow, and Cloud Storage to ingest and transform large-scale telecom datasets for analytics and reporting.
- Implemented Pub/Sub streaming ingestion to process event logs and network activity in near real time, improving operational visibility and data freshness.
- Optimized BigQuery SQL by applying partitioning, clustering, materialized views, and window functions to accelerate analytical queries.
- Built reusable Dataflow templates for both batch and streaming pipelines, improving maintainability and standardizing pipeline deployments.

- Automated data quality checks using schema validation, NULL profiling, and anomaly detection to ensure accuracy in downstream dashboards.
- Supported on-prem to GCP migration efforts by redesigning ingestion patterns to improve scalability and reduce infrastructure overhead.
- Collaborated with BI teams to deliver curated, analytics-ready datasets powering Looker Studio dashboards for telecom network insights.

Skills

- **Programming & Scripting:** Python, SQL, C, C++, Java, PySpark, Bash, Shell Scripting
- **AWS:** S3, Glue, Glue Crawler, Lambda, Step Functions, CloudWatch, SQS, SNS, IAM, EMR, Athena, Event Bridge, RDS
- **Azure:** ADF, ADLS Gen2, Synapse Analytics, Blob storage
- **GCP:** BigQuery, Dataflow, Pub/Sub, Dataproc, Cloud Storage, Looker Studio
- **Data Engineering & Big Data:** Spark, ETL/ELT Pipelines, Kafka, Airflow, Delta Lake, Lakehouse, Hadoop, Hive, HDFS, Data Quality Frameworks, Data & Dimensional Modeling (Star/Snowflake), Medallion Architecture, Databricks, Data Lakes
- **Databases:** PostgreSQL, MySQL, SQL Server, MongoDB, Snowflake, Redshift
- **Orchestration & CI/CD:** Airflow, Jenkins, Git, GitHub, Terraform, Docker, Kubernetes
- **Visualization & Analytics:** Power BI, Tableau, Looker Studio
- **Security & Governance:** IAM, Data Masking, Encryption, Schema Evolution, Lineage Tracking, Data Quality Validation, HIPAA-compliant handling of PHI/PII, Access Control (RBAC), Auditability & Metadata Logging

Project Work

Azure Lakehouse End-to-End Analytics (Azure) **Oct 2024 - Nov 2024**

Tech: ADF, ADLS, Databricks, Delta Lake, PySpark, Synapse, Power BI

- Built a Lakehouse using ADF, ADLS, Databricks, and Delta Lake to support scalable ingestion, schema evolution, and PySpark transformations for multi-format datasets.
- Modeled analytical views in Synapse with star schema and incremental loads, enabling fast Power BI reporting for research and analytics teams.

Real-Time Streaming Pipeline with Pub/Sub, Dataflow and BigQuery (GCP) **Feb 2024 - Mar 2024**

Tech: Pub/Sub, Dataflow, Apache Beam, BigQuery, Cloud Functions, Cloud Logging

- Created a real-time streaming pipeline with Pub/Sub for ingestion, Dataflow for event processing, and BigQuery for analytics-ready storage.
- Added schema validation, dead-letter handling, and Cloud Functions monitoring to ensure a reliable and observable streaming workflow.

Modern ELT Platform using Airflow, Snowflake and dbt (AWS and Snowflake) **Jun 2023 - Aug 2023**

Tech: Airflow, Snowflake, dbt, Snowpipe, S3, Python

- Built an Airflow-driven ELT system using Snowflake and dbt to automate ingestion, staging, and transformation workflows.
- Implemented Snowpipe ingestion and dbt models for staging and dimensional layers, supporting consistent and high-quality BI datasets.

Real-Time Healthcare Events Pipeline using Kafka, Spark Streaming and Snowflake (Kafka) **Jun 2023 - Aug 2023**

Tech: Kafka, Spark Structured Streaming, Snowflake, Snowpipe, Python, IAM, Encryption

- Built a real-time healthcare pipeline using Kafka for claims, eligibility, and clinical event ingestion and processed continuous streams in Spark Structured Streaming to create analytics-ready datasets.
- Implemented Snowpipe ingestion with schema checks, PHI and PII protections, and HIPAA-aligned access controls to support secure and reliable clinical and operational reporting.

Education

University of Arkansas at Little Rock | MS in Computer Science | Little Rock, Arkansas **Aug 2023 - May 2025**

RVRJC University | Bachelor's in Computer Science | Guntur, India **Jun 2019 - May 2023**