

Harsha Sai Ganesh Pabolu

Dayton, OH • (326) 235-7749 • pabuluharsha@gmail.com • [LinkedIn](#)

PROFESSIONAL SUMMARY

Senior Data Engineer with 5+ years building governed, audit-ready lakehouse platforms for payers and health systems at Fifth Third Bank, The Travelers, and Molina Healthcare. Delivered sub-15-minute claims reporting latency, 99.8% data freshness SLAs, HIPAA-compliant data products across 400+ datasets, and six-figure annual cloud cost reductions. Works across AWS and Azure with Apache Kafka, Spark, Delta Lake, Snowflake, Databricks, and dbt, and is completing a Google Cloud Professional Data Engineer certification to round out multi-cloud coverage.

TECHNICAL SKILLS

Languages & Libraries: Python, PySpark, SQL, Scala, Bash, T-SQL, SparkSQL, HiveQL, Pandas, NumPy, boto3, PyTest

Big Data & Streaming: Apache Spark, Spark Structured Streaming, Apache Kafka, Kafka Connect, Kafka Streams, Schema Registry, Debezium, CDC, Delta Lake, Delta Live Tables, Apache Hive, HDFS, Parquet, Avro, ORC, Lakehouse Architecture

Cloud Platforms: AWS (EMR, Glue, S3, Redshift, Kinesis, Lambda, Lake Formation, Athena, EventBridge, IAM, KMS) • Azure (Databricks, ADLS Gen2, Data Factory, Synapse Analytics, Event Hubs, Purview, Key Vault)

Data Warehousing & Modeling: Snowflake, Databricks SQL, Amazon Redshift, Azure Synapse Analytics, Delta Lake, Unity Catalog, Star Schema, Data Vault 2.0, Dimensional Modeling, Medallion Architecture, SCD, Z-ordering, Partitioning

Orchestration & ELT: Apache Airflow, dbt Core, dbt Cloud, Azure Data Factory, AWS Glue, dbt tests, dbt snapshots, data contracts, Fivetran

Data Quality & Governance: Great Expectations, Monte Carlo, Soda, Deequ, Microsoft Purview, Apache Atlas, OpenLineage, schema drift detection, SLA monitoring, PII masking, RBAC, HIPAA, GDPR, CCPA

DevOps, Observability & Tools: Terraform, Docker, Kubernetes, Helm, GitHub Actions, Jenkins, Azure DevOps, SonarQube, Datadog, Grafana, Azure Monitor, CI/CD, Agile, Scrum, JIRA, Git, PostgreSQL, MySQL, MongoDB, Oracle, Power BI

PROFESSIONAL EXPERIENCE

Senior Data Engineer | Fifth Third Bank | Cincinnati, OH

May 2024 - Present

- Architected end-to-end PySpark ETL pipelines on AWS EMR processing 2M+ daily insurance policy and claims records with exactly-once semantics, cutting batch processing time by 40% through partition and columnar storage tuning.
- Designed a bronze/silver/gold Delta Lake medallion architecture on AWS S3 consolidating 10 siloed data sources, enabling self-serve analytics for 200+ business users across actuarial, underwriting, and finance.
- Engineered real-time CDC ingestion with Apache Kafka, AWS Kinesis, and Debezium, reducing reporting lag from 24 hours to under 15 minutes with zero-downtime schema evolution across 12 producer services.
- Led adoption of dbt Core models with tests, snapshots, and SLA-backed data contracts, delivering governed data products to 6 downstream analytics teams at a 99.8% freshness SLA.
- Deployed Great Expectations and Monte Carlo observability across all gold-layer tables, authoring 100+ validation rules and cutting data incidents by 60% with proactive schema drift alerting.
- Implemented Microsoft Purview for lineage, cataloging, and PII classification across 400+ datasets, achieving full HIPAA compliance coverage and passing 3 consecutive audits.
- Automated infrastructure provisioning with Terraform and Helm on EKS, standardizing 8 pipeline environments and reducing deployment time from 4 hours to under 30 minutes.
- Optimized Delta Lake tables via Z-ordering, partitioning, and VACUUM scheduling, cutting Redshift query and compute costs by \$110K annually; mentored 3 engineers on PySpark and Delta Lake practices.

Data Engineer | The Travelers Inc. | Atlanta, GA

July 2022 - June 2023

- Delivered Azure Data Factory ELT pipelines processing 1.5M+ daily member, claims, and provider records from mainframe, Oracle, and Salesforce sources into Azure Synapse Analytics, improving analytical availability by 70%.
- Tuned Spark jobs via broadcast joins, partitioning, and columnar storage optimization on Azure Databricks, reducing average pipeline runtime by 43% and Azure compute spend by \$90K per year.
- Built Apache Kafka event streams on Azure Event Hubs with Debezium CDC connectors capturing real-time member changes, achieving sub-200ms alert latency with exactly-once semantics.
- Engineered Spark Structured Streaming jobs on Azure Databricks for claims risk scoring, processing 600K+ daily events with 99.9% SLA compliance.
- Architected a lakehouse on ADLS Gen2 using Delta Lake, consolidating 8 siloed data stores and reducing storage costs by 30%, and modeled star schema and data vault 2.0 structures in Synapse Analytics.
- Deployed Great Expectations and Monte Carlo monitors across claims pipelines, catching 92% of data quality issues before reaching the gold layer.
- Implemented HIPAA-compliant PII masking and column-level encryption using Azure Key Vault and Synapse row-level security, satisfying GDPR and internal governance policies.
- Provisioned 4 Azure environments with Terraform and Azure DevOps, authored Airflow DAGs replacing 25+ cron jobs, and established Jenkins CI/CD with pytest and dbt test gates at 82% code coverage.

Junior Data Engineer | Molina Healthcare | Long Beach, CA

March 2020 - June 2022

- Developed Azure Data Factory pipelines ingesting claims, member, and provider data from 6 source systems into Azure Synapse Analytics, enabling consolidated analytics for 400K+ member records.
- Built PySpark ETL jobs on Azure Databricks for claims adjudication data, reducing nightly batch processing time by 35% through partitioning and columnar storage optimization.
- Implemented HIPAA-compliant PII masking and encryption via Azure Key Vault, and applied Delta Lake on ADLS Gen2 for incremental member enrollment loads with full audit history and time-travel support.
- Built dbt models and SQL stored procedures for claims cost and utilization reporting, reducing analyst query time by 40% via pre-aggregated star schema tables in Synapse Analytics.
- Authored 80+ Great Expectations validation rules on claims pipelines, detecting and preventing 12+ critical data quality incidents before downstream impact.
- Contributed to PySpark pipelines for HEDIS and CMS Star Ratings regulatory reporting, processing 3M+ annual claims records across 22 quality measures with 100% on-time submission delivery.
- Built Apache Atlas metadata tagging and lineage mapping for 120+ regulatory datasets, supporting NCQA and CMS audit traceability requirements.
- Authored Apache Airflow DAGs for regulatory extract orchestration with retry logic and SLA monitoring, reducing pipeline failure rate by 45% across 14 compliance workflows.

PROJECTS

Real-Time Claims & Fraud Signal Streaming Platform - Fifth Third Bank (2025-Present) - Apache Kafka, AWS Kinesis, Spark Structured Streaming, Delta Lake, AWS EMR, Great Expectations, Terraform, Datadog

- Architected an event-driven streaming platform ingesting 2M+ daily claims and policy events via Apache Kafka and AWS Kinesis into a Spark Structured Streaming layer, cutting anomaly detection latency from hours to under 15 minutes.
- Built exactly-once ingestion with Schema Registry-enforced Avro contracts and Great Expectations quality gates feeding the Delta Lake gold layer, and instrumented Datadog dashboards that cut mean time to detection for pipeline failures.

Pharmacy & Care Management 360 Data Integration - The Travelers (2022-2023) - Azure Data Factory, Azure Databricks, Delta Live Tables, Azure Event Hubs, ADLS Gen2, dbt, MLflow, Great Expectations

- Built Azure Data Factory and Databricks pipelines integrating pharmacy claims, prior authorization, and care management data from 7 source systems into a unified 360-degree member view for care coordinators.
- Designed incremental Delta Live Tables pipelines with auto-loader for continuous ingestion and automatic schema evolution across 150GB+ of daily clinical data, and built MLflow-integrated feature pipelines serving 3 data science teams.

Enterprise Data Quality & Lineage Governance Framework - Fifth Third Bank (2025-Present) - Great Expectations, Monte Carlo, Microsoft Purview, OpenLineage, dbt tests, Python, Docker, GitHub Actions

- Designed a DataOps governance framework combining Great Expectations, Monte Carlo, and dbt tests with OpenLineage-based lineage tracking across all gold-layer tables, cutting undetected data incidents by 60%.
- Extended Microsoft Purview cataloging and PII classification into a self-serve compliance view for 6 downstream teams, standardizing HIPAA audit evidence collection and reducing manual audit prep effort.

Healthcare Claims Data Warehouse Migration - Molina Healthcare (2020-2022) - Azure Databricks, Azure Synapse Analytics, PySpark, dbt, Great Expectations, Terraform

- Led migration of legacy claims reporting from stored procedures to a PySpark and dbt-based pipeline on Azure Databricks, cutting nightly refresh time and improving maintainability for the analytics team.
- Partnered with actuarial and finance stakeholders to redesign star schema fact tables in Synapse Analytics, reducing report build time and standardizing claims cost metrics across teams.

EDUCATION

Master of Science in Computer Science - University of Dayton, Dayton, OH

Bachelor of Technology in Computer Science Engineering - Parul University, Vadodara, India

CERTIFICATIONS

- AWS Certified Data Engineer - Associate (DEA-C01)
- Databricks Certified Data Engineer Associate
- dbt Certified Analytics Engineer
- Google Cloud Professional Data Engineer (in progress, Target: Q3 2026)