Phaneendra Kumar Srungarapu

Dayton, OH • (937) 231-0449 • phaneendra.srungarapu1@gmail.com • LinkedIn

PROFESSIONAL SUMMARY

Senior Data Engineer with 5+ years building high-scale, low-latency, compliance-critical data pipelines across healthcare, insurance, and retail. Delivered 99.7% uptime, sub-200ms alert latency, zero-finding HIPAA/GDPR/CCPA audits, and $230K+ in annual cost savings at Target, Liberty Mutual, and Molina Healthcare. Trusted to architect and own enterprise data platforms end to end across AWS, Azure, and GCP using Apache Spark, Apache Kafka, Snowflake, Databricks, and dbt.

TECHNICAL SKILLS

Languages & Libraries: Python, PySpark, SQL, Scala, Bash, T-SQL, Pandas, NumPy, SQL Server
Big Data & Streaming: Apache Spark, Spark Structured Streaming, Apache Kafka, Kafka Streams, Kafka Connect, Apache Flink, Schema Registry, Debezium, CDC, Delta Lake, Apache Iceberg, Delta Live Tables, Lakehouse Architecture
Cloud Platforms: AWS (S3, Glue, EMR, Redshift, Kinesis, Athena, Lake Formation, MWAA, SageMaker, Lambda, EKS) • Azure (Databricks, Synapse Analytics, ADLS Gen2, Event Hubs, ADF, Purview, Key Vault, AKS, Monitor) • GCP (BigQuery, Dataflow, Dataproc, Pub/Sub, Composer, Vertex AI)
Data Warehousing & Modeling: Snowflake, Databricks SQL, Azure Synapse Analytics, Amazon Redshift, BigQuery, Unity Catalog; Star Schema, Data Vault 2.0, Dimensional Modeling, Medallion Architecture, Slowly Changing Dimensions, Data Mesh
Orchestration & ELT: Apache Airflow, dbt Core, dbt Cloud, dbt Mesh, dbt Semantic Layer, Azure Data Factory, Dagster, Prefect, Fivetran, Airbyte
Data Quality, Governance & DevOps: Great Expectations, Monte Carlo, Soda, dbt tests, Purview, Atlas, OpenLineage, Data Contracts, Anomaly Detection, PII Masking, RBAC, MLflow, Feature Engineering, Feast, Docker, Kubernetes, Terraform, Helm, GitHub Actions, CI/CD, Azure DevOps, Datadog, OpenTelemetry, HIPAA, GDPR, CCPA

PROFESSIONAL EXPERIENCE

Senior Data Engineer | **Target Corporation** | Minneapolis, MN *August 2024 – Present*

- Architected Apache Kafka and Azure Event Hubs real-time ingestion layer on Azure Databricks processing 3M+ daily retail transactions with exactly-once semantics and sub-500ms end-to-end latency.
- Designed Delta Lake medallion architecture (bronze, silver, gold) on ADLS Gen2, reducing data redundancy by 45% and enabling self-serve analytics for 200+ business users.
- Built Azure Data Factory ELT pipelines ingesting supply chain, POS, and vendor data from 15+ heterogeneous source systems into Azure Synapse Analytics, cutting load times by 38%.
- Developed dbt models with tests, snapshots, and dbt Mesh domain isolation across inventory, pricing, and fulfillment domains, achieving 99.7% data freshness SLA.
- Deployed Great Expectations and Monte Carlo data observability across all gold-layer tables, reducing data incidents by 62% and enabling proactive schema drift alerting.
- Implemented Microsoft Purview for end-to-end data lineage, catalog, and PII classification across 500+ datasets, achieving full CCPA compliance.
- Provisioned infrastructure with Terraform and Helm on AKS, standardizing 12 pipeline environments and reducing deployment time from 4 hours to 22 minutes.
- Optimized Delta Lake tables via Z-ordering, partitioning, and VACUUM scheduling, cutting Azure Synapse query costs by $120K annually.

Data Engineer | **Liberty Mutual Insurance** | Boston, MA *February 2021 – July 2023*

- Delivered Azure Data Factory ELT pipelines processing 2M+ daily insurance policy, claims, and customer records from mainframe, Oracle, and Salesforce into Azure Synapse Analytics, improving analytical availability by 70%.
- Built Apache Kafka event streams on Azure Event Hubs with Debezium CDC connectors capturing real-time policy changes, reducing reporting latency from 24 hours to under 15 minutes.
- Architected data lakehouse on ADLS Gen2 using Delta Lake and Apache Iceberg, consolidating 8 siloed data stores and reducing storage costs by 32%.
- Designed Kafka Streams topology for real-time fraud signal detection across 500K+ daily transactions, achieving sub-200ms alert latency with exactly-once semantics.
- Built Spark Structured Streaming pipelines on Databricks feeding ML fraud models via Feast feature store, reducing false-positive fraud rate by 18% through MLflow A/B testing.
- Enforced HIPAA-compliant PII masking via Azure Key Vault and column-level security in Synapse Analytics, passing 3 consecutive compliance audits with zero findings.
- Authored Apache Airflow DAGs replacing 40+ cron jobs for multi-dependency orchestration, reducing pipeline failures by 67% and SLA breach incidents by 80%.
- Optimized Spark jobs via broadcast joins, dynamic partitioning, and columnar storage tuning, reducing pipeline runtime by 43% and Azure compute spend by $90K annually.

Data Engineer | **Molina Healthcare** | Long Beach, CA *February 2020 – January 2021*

- Built Azure Data Factory pipelines ingesting claims, member, and provider data from 6 source systems into Azure Synapse Analytics, enabling consolidated analytics for 500K+ member records.
- Developed PySpark ETL jobs on Azure Databricks for claims adjudication, reducing nightly batch processing time by 35% through dynamic partitioning and columnar storage optimization.
- Implemented HIPAA-compliant PII masking and encryption via Azure Key Vault, protecting PHI across all analytical environments for 500K+ members; optimized Spark reducing compute spend by $20K annually.
- 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 data quality validation rules on claims pipelines, detecting and preventing 15+ critical data incidents before reaching production.
- Contributed to PySpark pipelines for HEDIS and CMS Star Ratings regulatory reporting, processing 3M+ annual claims records with 100% on-time government submission delivery.
- Implemented Datadog and Azure Monitor observability across 12 production pipelines, reducing mean time to detection by 55% and enabling proactive SLA breach alerting.
- Collaborated with data science and actuarial teams to build feature engineering pipelines on Azure Databricks, improving population health model accuracy by 22%.

PROJECTS

Real-Time Market Analytics Pipeline | Target (2024-2025) | Python, Apache Kafka, PySpark, Spark Structured Streaming, Delta Lake, Snowflake, dbt, Apache Airflow, Terraform, AWS, Grafana, GitHub Actions

- Architected end-to-end streaming pipeline ingesting 500K+ retail transaction events/min via Apache Kafka → Spark Structured Streaming → Delta Lake medallion architecture → Snowflake, achieving sub-600ms latency with 99.9% uptime.
- Implemented dbt transformation layer with 150+ automated data quality tests, Schema Registry-enforced Avro schemas, and exactly-once semantics; integrated Grafana and OpenTelemetry dashboards reducing pipeline failure MTTD by 70%.

Automated Data Quality & Governance Framework | Liberty Mutual (2022-2023) | Python, dbt, Great Expectations, Monte Carlo, OpenLineage, Snowflake, Apache Airflow, FastAPI, Streamlit, Docker, Kubernetes, GitHub Actions

- Built open-source DataOps framework integrating Great Expectations, Monte Carlo, and dbt tests with OpenLineage for automated end-to-end data quality governance across multi-cloud pipelines spanning Snowflake and Databricks.
- Developed FastAPI microservice and Streamlit observability dashboard surfacing 300+ validation rules, anomaly detection alerts, and real-time data lineage graphs; containerized with Docker and Kubernetes, deployed via CI/CD.

Cloud-Native Healthcare Data Lakehouse | Molina Healthcare (2020-2021) | PySpark, Delta Lake, Apache Iceberg, AWS Glue, Amazon Redshift, dbt, Great Expectations, Apache Airflow, Terraform, AWS Lake Formation, SageMaker

- Designed multi-cloud healthcare data lakehouse on AWS using Apache Iceberg over S3 with medallion architecture, ingesting 5M+ daily records from HL7/FHIR APIs and EHR systems into Amazon Redshift via AWS Glue and Lake Formation.
- Implemented HIPAA-compliant PII tokenization, column-level encryption, and SageMaker-integrated feature pipelines for clinical risk models; deployed full IaC via Terraform with GitHub Actions CI/CD achieving 99.95% pipeline SLA.

EDUCATION

Master of Science in Computer Science — University of Dayton, Dayton, OH
Bachelor of Technology in Computer Science and Engineering — Parul University, Vadodara, India

CERTIFICATIONS

- AWS Certified Data Engineer – Associate (DEA-C01)
- Google Cloud Certified Professional Data Engineer
- Databricks Certified Data Engineer Associate
- Snowflake SnowPro Core Certified
- dbt Certified Analytics Engineer