CHAITANYA PRIYA SURUKONTI
Senior Senior Data Engineer

SUMMARY

- Senior Data Engineer with 5+ years designing scalable lakehouse architectures and distributed data pipelines using Databricks, Snowflake (Snowpark), PySpark, Kafka, and Airflow across healthcare, life sciences, and finance domains.
- Expertise in building AI/ML-ready feature engineering pipelines, embedding datasets, and integrating ML workflows with MLflow and Databricks Feature Store for clinical risk modeling and forecasting.
- Proficient in cloud platforms AWS, Azure, and GCP, implementing CI/CD, Terraform IaC, and DataOps frameworks ensuring HIPAA-compliant governance and 99.5% SLA data freshness.
- Strong background in performance optimization, including Snowflake query tuning (28% improvement), Spark resource tuning (18% cost reduction), and streaming ingestion reducing latency by 45%.
- Master of Science in Computer Science with hands-on experience in REST API development (FastAPI), containerization (Docker, Kubernetes), and modular data contracts using dbt, enabling robust, scalable data engineering solutions.

TECHNICAL SKILLS

Main Technical Skills	Python (5 yr.), PySpark (5 yr.), Apache Spark (5 yr.), SQL (5 yr.), Databricks (3 yr.)
Programming Languages	Python (5 yr.)
Java Frameworks	Apache Spark (5 yr.)
Scala Frameworks	Apache Spark (5 yr.)
Python Frameworks	FastAPI (3 yr.)
AI & Machine Learning	Mlflow (3 yr.), Vertex AI
Python Libraries and Tools	PySpark (5 yr.)
Data Analysis and Visualization Technologies	Apache Airflow (4 yr.), Apache Spark (5 yr.), Apache Spark Streaming (3 yr.), Databricks (3 yr.), Looker Studio, Power BI (2 yr.), Tableau (2 yr.)
Databases & Management Systems / ORM	Apache Spark (5 yr.), Apache Spark Streaming (3 yr.), AWS Redshift, dbt (3 yr.), Oracle Database (2 yr.), PostgreSQL (4 yr.), Snowflake (3 yr.), SQL (5 yr.)
Cloud Platforms, Services & Computing	AWS (4 yr.), GCP
Amazon Web Services	AWS Lambda (2 yr.), AWS Redshift
Azure Cloud Services	Databricks (3 yr.)

Google Cloud Platform	Google BigQuery
UI/UX/Wireframing	3D Modelling (5 yr.)
Deployment, CI/CD & Administration	CI/CD (3 yr.)
QA, Test Automation, Security	Data Validation (4 yr.)
Virtualization, Containers and Orchestration	Docker (3 yr.), Kubernetes (3 yr.), Terraform (3 yr.)
SDK / API and Integrations	FastAPI (3 yr.), RESTful API (3 yr.)
Message/Queue/Task Brokers	Kafka (3 yr.)
Methodologies, Paradigms and Patterns	Publish/Subscribe Architectural Pattern
Other Technical Skills	DataOps (3 yr.), Delta lake (3 yr.), Snowpark API (3 yr.), Spark EMR (4 yr.)

WORK EXPERIENCE

Senior Data Engineer - CVS Health (Healthcare Data Lakehouse and AI-Ready Pipelines)

Duration: Jun 2025 – Present
Summary:

- Development and architecture of scalable batch and near-real-time data pipelines processing multi-terabyte healthcare datasets daily
- The project supports clinical risk modeling, operational forecasting, and AI retrieval workflows by delivering ML-ready and embedding-ready datasets
- It includes integration of ML workflows and deployment automation across cloud environments

Responsibilities:

- Architected scalable data pipelines using PySpark, Databricks, Delta Live Tables, and Kafka.
- Built AI-ready feature engineering pipelines for ML training and inference.
- Designed and optimized Snowflake pipelines with Snowpark, Streams & Tasks.
- Implemented ingestion pipelines for structured and semi-structured healthcare data supporting AI retrieval workflows.
- Developed embedding-ready datasets for AI experimentation.
- Integrated ML workflows with MLflow, Databricks Feature Store, and model versioning.
- Designed REST-based ingestion services using FastAPI.
- Orchestrated pipelines using Apache Airflow and Terraform for deployment automation on AWS and Azure.
- Implemented dbt transformation layers for modular data contracts and semantic models.
- Built DataOps validation frameworks for schema drift detection, anomaly checks, and observability monitoring.
- Enforced HIPAA-compliant governance with RBAC policies and audit-ready lineage.

- Optimized compute costs via Spark resource tuning and cluster auto-scaling.

Technologies: Databricks, Delta Lake, Snowflake (Snowpark, Streams, Tasks), PySpark, Kafka, dbt, MLflow, FastAPI, AWS (S3, Glue, EMR), Azure Databricks, Terraform, Docker, Kubernetes

Data Engineer – ML & Analytics - Dr. Reddy's Laboratories (Supply Chain and Forecasting Data Pipelines)

Duration: Mar 2021 – Jul 2023
Summary:

- Designed and implemented Spark-based ELT pipelines supporting enterprise analytics and machine learning initiatives across supply chain and forecasting systems
- Migrated on-premises workflows to AWS S3 Data Lake architecture to improve scalability and reduce costs
- Developed AI-ready datasets and optimized Snowflake performance for secure data sharing and efficient transformations

Responsibilities:

- Designed Spark-based ELT pipelines for analytics and ML initiatives.
- Migrated on-prem workflows to AWS S3 Data Lake architecture.
- Built scalable feature engineering pipelines for ML training.
- Designed curated AI-ready datasets using dimensional modeling and Snowflake transformations.
- Implemented Snowflake performance tuning, partition optimization, and secure data sharing.
- Developed data ingestion workflows from REST APIs and third-party sources.
- Integrated batch and streaming ingestion using Kafka to reduce reporting latency.
- Implemented data validation pipelines with profiling and reconciliation logic.
- Orchestrated workflows with Apache Airflow maintaining high data freshness SLAs.
- Collaborated with data scientists on feature refresh schedules and schema evolution for model retraining.

Technologies: AWS (S3, EMR, Lambda), Apache Spark, PySpark, Snowflake, Airflow, PostgreSQL, Tableau, Python, Kafka

Data Engineer - Hexaware Technologies (Financial Data ETL and Fraud Analytics Support)

Duration: Mar 2019 – Feb 2021
Summary:

- Developed SQL-based ETL pipelines ingesting financial datasets for risk and fraud analytics in regulatory environments
- Automated data cleansing workflows and designed relational data models to support reporting and risk analysis
- Delivered curated datasets for predictive risk scoring and built dashboards tracking operational KPIs and compliance metrics

Responsibilities:

- Developed SQL-based ETL pipelines for financial data ingestion.
- Automated data cleansing workflows using Python.
- Designed relational data models in Oracle and PostgreSQL.
- Supported fraud analytics teams with curated datasets.
- Implemented validation and reconciliation checks for data accuracy.

- Optimized SQL queries and ETL jobs to improve reporting performance.
- Built Power BI dashboards for operational KPIs and compliance metrics.

Technologies: SQL, Python, Oracle, PostgreSQL, Power BI

EDUCATION

- **Master of Science in Computer Science**
  University of Bridgeport — Bridgeport, Connecticut, USA
  Sept 2023 - May 2025