Sivani Ganti

(443) 857-3385 | sivaniganti16@gmail.com | linkedin.com/in/SivaniGanti

SUMMARY

Data Engineer with **4+ years** of expertise designing and implementing large-scale data pipelines across **AWS and Azure** using **Python, SQL, PySpark**, and **Kafka**. Proficient in building secure, production-grade **batch and streaming** ETL/ELT workflows for transactional systems, enabling analytics, reporting, and ML use cases. Skilled in **data modeling** (star/snowflake schemas, SCD/CDC), **data lineage, and governance, performance tuning**. Experienced with **Airflow** orchestration and modular ELT development using **dbt, Databricks, Glue**. Architected scalable **Snowflake** environments and implemented **CI/CD** automation using GitHub Actions, **Jenkins**, **Docker**, **Terraform**, and **Ansible**. Adept at collaborating across data science, analytics, ML engineering teams in **Agile** environments to deliver resilient, SLA-driven data platform.

EXPERIENCE

HD Supply (Home Depot) St.Louis, MO

Cloud Data Engineer *[PySpark, Kafka Streams, SQL, Airflow, AWS, EMR, Glue, Lambda, dbt, Snowflake]* *May 2024 – Present*

- Led the migration of **20TB** of warehouse inventory transactions and order fulfillment logs data from on-prem legacy systems, **Oracle** (via **AWS DMS** with **CDC**), **Hadoop/Hive** (via **DistCp)**, to an **AWS Data Lake**, building scalable batch pipelines with **PySpark** on **Glue/EMR** and integrated **Kafka** for real-time ingestion. Loaded curated datasets into **Snowflake**, cutting infrastructure costs by **40%**, and enabling centralized analytics across business teams.
- Applied **Medallion architecture** on **S3** with partitioned Parquet outputs and **Glue Catalog registration,** enabling **Athena** queries and Snowflake ingestion. Modeled Snowflake datasets using **dbt** staging and mart layers with **SCD Type 2 logic**, and leveraged **external tables and data sharing** to enable governed analytics across product, supply chain, and ML workflows.
- Designed and implemented real-time data pipelines using **Kafka Streams** and **PySpark Structured Streaming on EMR** to process **30M+** daily warehouse events including inventory updates, worker activity logs, order picking and packing actions.Utilized producer-consumer patterns with partitioning, retry logic, and Dead Letter Queues (DLQ) to ensure high throughput and fault tolerance.
- Streamed structured events into Snowflake for real-time analytics, and parsed **nested JSON** payloads using VARIANT columns and FLATTEN functions to build analytics-ready views. Ensured **data integrity** by developing **automated validation** using Python/Pyspark scripts to detect and report inconsistencies, schema mismatches, data drift in Snowflake. Leveraged Snowflake's INFORMATION_SCHEMA for row-level reconciliation between curated S3 data and Snowflake tables.
- Architected normalized **ER models** in **Erwin** with SCD Type 2 and CDC logic, designing **star/snowflake schemas** to support historical tracking and scalable, cross-domain data pipelines. Optimized **Snowflake** data models using **schema versioning**, **RBAC**, **clustering keys, time travel, and materialized views**, reducing compute costs by **35%**.
- Refactored **complex SQL transformations** and optimized stored procedures in Snowflake to support multi-step analytics workflows and compliance dashboards in **Tableau**, reducing downstream query latency by **45%** and accelerating data access for business users.
- Orchestrated end-to-end data pipelines using **Airflow**, leveraging AWS **Lambda** for event-driven triggers and **Ansible** for automated EMR and DAG provisioning in production environments. Built **CI/CD** pipelines using **Git**, **Jenkins**, and **Docker containerization** to enable production-grade deployments across (dev-to-prod), standardizing Git branching strategies for release **governance** and platform reliability
- Developed reusable **PySpark scaffolds** and **Pytest**-based validation utilities to enforce schema checks, improve pipeline reliability, and ensure end-to-end **data lineage**. Standardized development patterns to accelerate onboarding and enhancing collaboration across engineering teams.
- Provisioned Snowflake infrastructure using **Terraform** and **AWS CLI**, managing **IAM roles** and **Secrets Manager credentia**ls to securely connect AWS-based pipelines and enable reproducible deployments across environments.
- Integrated **DynamoDB** for dynamic product attributes and category filters to enable personalized search, with **PostgreSQL** supporting low-latency API lookups alongside **Elasticsearch**-based semantic search.
- Developed **Python scripts** to ingest third-party logistics, SLA, and SKU pricing data via **APIs**, normalize it, and prepare curated datasets for Airflow-triggered ingestion into Snowflake, expanding external data coverage for supply chain analytics.
- Set up SLA monitoring and **CloudWatch alerts** on Glue/EMR jobs to detect early issues in Snowflake data loads, and integrated **SNS notifications** to **trigger Slack alerts** for job failures and late task executions.

Goldman Sachs New York City, NY

Software Engineer - Data *[SQL, Python, Spark, S3, AWS MSK, CloudFormation, Redshift, Prometheus]* *October 2023 – March 2024*

- Migrated high-volume **Kafka** workloads from Standard MSK to **MSK Express**, using MSK Replicator for seamless topic and offset transfer and **CloudFormation** for automated cluster provisioning, boosted throughput **3×** and lowered latency with zero production impact.
- Built **Spark SQL** and **Redshift** pipelines to aggregate Kafka-ingested event streams and supplied feature data for

fraud and bot detection models, enabling near **real-time alerting** and reducing false positives by 35%.

- Modeled transactional and behavioral datasets in Redshift with SCD logic and star schema to support scalable reporting and dynamic risk classifications with full auditability, reducing manual- investigation by 45%
- Refactored **batch ETL jobs** into **Spark Structured Streaming** pipelines using **AWS EMR** with **S3-backed checkpoints,** AWS **Glue Data Catalog** for centralized schema, enabling sub 10s latency on critical data flows such as real-time transaction monitoring, suspicious activity alerts, and fraud scoring triggers and compliance alerts.
- Rewrote core fraud scoring pipeline in **Java**, integrated with **AWS Lambda**. Reduced daily ETL runtime from 6 hours to 45 minutes, enabling real-time alerts for 10M+ transactions.
- Optimized **Spark** streaming jobs with **adaptive memory tuning and dynamic partitioning**, resulting in 40% cost savings on **EC2** clusters and 2× faster processing of high-volume event streams (~ 50K+ msg's/sec).
- Optimized Kafka-to-Redshift integration by implementing Spark **micro-batching** with intermediate S3 staging, significantly improving data freshness and reliability for compliance reporting and downstream analytics.
- Built cross-region **Kafka replication pipelines** with **Prometheus and Grafana** monitoring and automated geo-failover, reducing disaster recovery time by 70% and ensuring uninterrupted data flow for 24/7 trading systems.
- Designed resilient streaming pipelines with auto-failover and **SLA-based alerting**, powering real-time **Tableau** dashboards for regulatory compliance, operational metrics, and behavioral analytics.
- Created modular **PySpark** job templates and parameterized **Airflow DAGs**, and collaborated with **ML teams** to deliver Redshift-ready datasets for anomaly detection and risk classification models. Developed reusable **Python utilities (UDFs, config generators, log parsers)** to standardize ingestion and downstream analytics workflows.

Snapdeal Hyderabad, India
Big Data Engineer *[SQL,Azure Databricks,Kafka,Azure Data Lake, Synapse,Graphana]* *February 2020 – December 2022*

- Implemented high-throughput ETL pipelines on **Databricks** to process over **100 GB/day** of user **telemetry data** from **Azure Data Lake**, enabling near-real-time user behavior analytics (15 min latency) for downstream teams.
- Leveraged **Cassandra** for sub-second lookups in real-time customer segmentation and recommendation pipelines supporting ML models and engagement dashboards.
- Built schema drift detection logic in **PySpark** to validate telemetry structures before ingestion, improving pipeline reliability and preventing silent failures in curated datasets.
- Developed an image optimization pipeline to auto-compress and resize product images, storing outputs in **Azure Blob Storage**. Improved page load times by **30%** and reduced storage and **CDN** costs by **40%**.
- Parameterized transformation logic in **notebook** to enable reuse across file formats and size thresholds for various image categories.
- Optimized Spark workflows using **RDD** and **DataFrame APIs** by **caching** intermediate transformations and reducing shuffle operations, achieving sub-second latency in user clickstream processing for real-time analytics and alerting.
- Tuned executor sizing, **broadcast joins**, and spill thresholds to reduce **Spark job runtime by 35%** and improve concurrency across batch and streaming loads.
- Automated end-to-end workflows using Azure Data Factory (**ADF**) to orchestrate daily clickstream aggregation, product catalog ETL, and **Kafka** topic compaction/cleanup. Monitored SLA-driven pipelines using **Azure Monitor alerts and Log Analytics queries**, raising job success rate from **~89% to over 98%**.
- Integrated ADF with **Git**-based deployment for version control and environment promotion, supporting rollback and auditability across dev, QA, and production pipelines.
- Delivered external tables and Power BI optimized **Synapse** views backed by pre-aggregated datasets, improving dashboard load times by 40% and enabling scalable self-service analytics even under peak load.

EDUCATION

Master's in Health and Information Technology Jan 2023 – Dec 2024
University of Maryland, Baltimore County, Baltimore, MD *GPA: **4.0/4.0***

SKILLS

Programming Languages & Frameworks: Python, SQL, PL/SQL, PySpark, Spark SQL, Java, Shell Scripting, YAML, Jinja, Pytest

Big Data & Distributed Computing: Apache Spark, Spark Structured Streaming, Kafka, Databricks, Hadoop (HDFS, Hive), EMR

Cloud Platforms & Services: AWS (Glue, EMR, Lambda, Athena, S3, Redshift, IAM, EKS, CloudWatch), Azure (Data Factory, Data Lake, Synapse), GCP (BigQuery)

Data Warehousing & Modeling: Snowflake, Amazon Redshift, Azure Synapse, dbt, SSIS, Dimensional Modeling (Star & Snowflake Schemas), SCD Types 1 & 2, CDC

ETL, Workflow Orchestration & Automation: Apache Airflow, AWS Glue, ADF, Databricks, Jenkins, Ansible, Docker, Terraform, GitHub Actions, CI/CD Pipelines

Databases: RDBMS(PostgreSQL, MySQL), NoSQL (MongoDB, DynamoDB,Cassandra), Neo4j.

Monitoring & Observability: Prometheus, Grafana, AWS CloudWatch

Analytics & Visualization Tools: Tableau, Power BI, Quicksight, Cloudwatch, Jupyter Notebook, Anaconda

Application Stacks & APIs: HTML5, CSS3, React.js, REST API

Development & Version Control Tools:VS Code, PyCharm, Git, GitHub, JIRA, Docker