Mohith Kasturi

mohithk.data@gmail.com • (331) 303-6476 • linkedin.com/in/-mohit-b219b7296 • github.com/mohithk

Professional Summary

Data Engineer with 5+ years of experience designing cloud-native systems that process millions of records daily across AWS, Azure, and GCP. Specialized in building event-driven, microservices-based real-time and batch pipelines using Python, SQL, Spark, dbt, and Snowflake to support analytics, machine learning, and regulatory reporting. Experienced in collaborating with engineering, product, and compliance teams in Agile/SAFe environments to reduce latency, improve pipeline reliability, and deliver scalable data solutions using orchestration and governance tools such as Airflow, dbt, and Amundsen across finance and healthcare domains.

Education

M.S. in Computer Science — **Governors State University**, Chicago, IL
GPA: 3.9/4.0 — Jan 2022 – Dec 2023

B.E. in Electrical Engineering — JNTU, Hyderabad, India

Technical Skills

- **Programming:** Python (Pandas, PySpark, pytest, unittest), PL/SQL, Java, Bash, Scala
- **Cloud & Big Data:** AWS (S3, Redshift, Glue, Lambda), Azure (ADF, Synapse, Event Hubs), GCP (BigQuery, Dataflow, Pub/Sub), Databricks, Hadoop (Hive, HDFS, Pig, Sqoop, MapReduce, Oozie, Zookeeper), Ab Initio (GDE, Metadata Hub), Snowflake
- **ETL/ELT & Orchestration:** Apache Spark, dbt, Apache Airflow, SSIS, Talend, SnowPipe
- **Storage & Warehousing:** Redshift, Synapse, BigQuery, PostgreSQL, Delta Lake, Hive, HDFS
- **Streaming & APIs:** Kafka, Kinesis, REST, GraphQL, MQTT, Event Hubs
- **Data Visualization:** Power BI, Tableau, Looker, Excel
- **DevOps & IaC:** Docker, Kubernetes, Terraform, Jenkins, GitHub Actions
- **Governance & Security:** RBAC, RLS, PII masking, Amundsen, DataHub, encryption, audit logging
- **Project Tools:** Jira, Confluence, Azure Boards, Agile/Scrum

Professional Experience

Azure Data Engineer, First Horizon Bank — Germantown, TN
Apr 2023 – Present

- Led scalable Azure and Snowflake data pipelines for regulatory reporting and fraud detection, reducing latency by 25%.
- Built Java API for real-time Cassandra access integrated with Hadoop, Solr, PySpark, Kafka, and Storm.
- Developed event-driven fraud detection workflows using Event Hubs, Databricks, Spark ML (XGBoost, scikit-learn).
- Optimized batch ETL from PostgreSQL, MongoDB, SQL Server, improving throughput by 40%.
- Automated infrastructure provisioning with Docker, Kubernetes, and Terraform, halving deployment time.
- Integrated real-time APIs and Kafka streams into Databricks, enhancing fraud feature generation.
- Refactored PySpark pipelines with dbt test coverage and pytest, reducing processing time by 30%.
- Implemented RBAC, audit logging, and Section 19 privacy controls in Synapse and Databricks.
- Collaborated with cross-functional teams to define data requirements for credit and fraud analytics.

AWS Data Engineer, Optum — Eden Prairie, MN
Apr 2022 – Mar 2023

- Engineered and optimized multi-terabyte ETL pipelines on AWS and Databricks, supporting 25K+ Medicare claims daily.
- Built secure ETL workflows with IAM, encryption, and audit logging for HIPAA compliance.
- Built pipelines ingesting 1M+ records daily from JSON, XML, and REST APIs via Glue, Redshift, Lambda.
- Developed cross-cloud workflows with PySpark integrating BigQuery, Snowflake, Redshift.

- Tuned Databricks Spark clusters for real-time transformations via Spark DataFrame API.
- Used Ab Initio (GDE, Metadata Hub) to manage metadata-driven workflows and orchestration.
- Refactored legacy Informatica ETLs into Spark and Snowflake, improving speed by 40%.
- Reverse-engineered SQL/C APIs and integrated logic into dbt and PySpark with unit testing.
- Delivered real-time analytics via SSIS, Spark, Snowflake, cutting dashboard load time by 50%.
- Built PySpark-based monitoring pipelines reducing incident response time.
- Created Tableau dashboards transforming risk metrics into actionable visuals.

GCP Data Engineer, Takeda Pharmaceutical Mumbai, India
Jan 2020 – Dec 2021

- Built real-time and batch ELT pipelines using Apache Beam, Dataflow, BigQuery for clinical analytics.
- Migrated legacy pipelines to GCP with PySpark, Hive, and DataFrame APIs.
- Designed ingestion for structured/semi-structured data across on-prem and cloud.
- Used Hadoop tools (Hive, HDFS, Pig, Sqoop, MapReduce, Oozie, Zookeeper) for scalable ETL.
- Implemented monitoring pipelines with Airflow, Cloud Functions, and Python.
- Orchestrated vitals ingestion using Pub/Sub, Event Hubs, and visualized via Power BI.
- Developed Git-driven Airflow DAGs deployed via Terraform and Azure DevOps CI/CD.
- Standardized schemas across OLAP/OLTP systems with schema validation for Power BI.
- Enforced secure and repeatable deployments with CI/CD and Terraform under DevSecOps.

Data Engineer, Amazon Mumbai, India
Jul 2019 – Dec 2019

- Developed predictive models using scikit-learn, XGBoost to forecast customer churn and risk.
- Built ELT pipelines with Glue, Lambda, Redshift integrating third-party customer data.
- Engineered cloud-native data lakes with BigQuery, Snowflake, Redshift, and S3.
- Analyzed 2.5M Prime member records, improving retention by 15%, revenue by $15M/month.
- Optimized Kafka-S3 pipelines and Redshift/Hive queries for better performance.
- Applied Jenkins and Terraform for CI/CD and infrastructure automation.
- Conducted ETL performance testing with JMeter, cutting latency by 20%.
- Streamed real-time behavior via Kinesis and EMR to drive targeted marketing.
- Defined data quality rules and delivered Power BI dashboards on customer churn.

Certifications

- Microsoft Azure Fundamentals (AZ-900)
- Google Cloud Practitioner
- AWS Certified Cloud Developer Associate
- Databricks Certified Data Engineer Associate (in progress)