# Vishal Sreepada

linkedin.com/in/vishal-sreepada/ • 571.234.1842
vishal.sreepada.77@gmail.com • vishalsreepada.com

## Summary

Data Engineer II at Amazon with 5+ years of experience building large-scale data lake infrastructure, event-driven pipelines, and AI-powered automation systems. Designed and owned a 63+ pipeline medallion architecture data lake serving global operations, built AI agents on Amazon Bedrock that automate metadata enrichment and pipeline triage, and led cross-org platform migrations at scale. Deep expertise in AWS-native data infrastructure, PySpark, and distributed systems.

## Education

**Master of Science in Computer Engineering** — May 2021
George Mason University, Fairfax, Virginia
*Outstanding Academic Achievement Award*

**Bachelor of Engineering in Electronics and Communication Engineering** — May 2019
Osmania University, Hyderabad, India

## Technical Proficiencies

| | |
|---|---|
| Languages: | Python, SQL, PySpark, Scala |
| Cloud Technologies: | Azure, AWS, Databricks |
| Databases: | Redshift, Aurora MySQL, DynamoDB, DocumentDB, Snowflake, S3, Apache Iceberg |
| AWS Services: | Glue, EMR Serverless, Lambda, Athena, Lake Formation, DataZone, QuickSight, EventBridge, SNS, SQS, SES, CloudFront, API Gateway, CodePipeline, CloudWatch |
| Azure Services: | Databricks, Storage accounts, Key vaults, App Services, Virtual networks |
| AI & Agents | Amazon Bedrock, Bedrock AgentCore, Strands Agents SDK, Langfuse |
| Tools: | Jupyter Notebook, Git, Jira, Visual Studio Code |

## Professional Experience

### Amazon

**Data Engineer II - Austin, TX** — May 2024 – Present

- Owned the enterprise data lake platform powering 63+ production pipelines, built on a 4-layer medallion architecture (Landing → Raw → Transform → Refine) with automatic compute selection across Lambda, AWS Glue, and EMR Serverless, and Apache Iceberg for ACID compliance on high-volume datasets.
- Automated registration and AI enrichment of 6,000+ datasets into the enterprise data catalog, replacing a 30-min-per-table manual process with a Strands Agent on Bedrock AgentCore that auto-generates metadata, descriptions, and data classifications at scale.
- Built an AI agent deployed on an internal agent platform that triggers on pipeline failure tickets, analyzes root cause using domain knowledge, and posts a detailed diagnosis comment directly on the ticket, reducing mean time to resolution from 4-5 days to 1-2 days.
- Reduced pipeline onboarding time from half a day to 30 minutes with a conversational AI interface where teams configure and launch new data sources through chat. Replaced a fixed weekly release cycle with instant production deployments, enabling same-day pipeline launches.
- Led org-wide migration to a next-generation data catalog platform, owning documentation and cross-team coordination to ensure zero disruption across the Engineering & Analytics organization.
- Automated analytics access provisioning by building a system that syncs identity groups to QuickSight permissions in near real-time, coordinating across LDAP, IAM, and Lake Formation to grant analysts instant access to sensitive datasets when authorized.

**Data Engineer - Seattle, WA** — Apr 2022 – Apr 2024

- Eliminated ~200 daily manual email reviews by building an event-driven ticket routing system that auto-routes warehouse sensor failures, deduplicates multi-device alerts, and auto-resolves transient issues.
- Built multi-vendor fleet telemetry pipelines ingesting data from 40+ API endpoints across US and European fulfillment centers, completing a 2-year historical backfill into the enterprise data lake while retiring a legacy Redshift cluster.
- Expanded the data lake to handle sensitive and restricted datasets by completing a formal security certification process, authoring architecture documentation and threat mitigations through cross-functional review.

- Reduced an executive analytics pipeline runtime by 98% (from 40+ hours to 30 minutes) by identifying 990GB of unnecessary data processing and redesigning with partition pruning and targeted S3 crawlers, unblocking weekly reporting for multiple business units.
- Migrated 600+ users across Europe off 9 legacy tools to a unified data platform in a planned one-day cutover with zero data loss, owning schema mapping for 50+ tables and all downstream compatibility pipelines.

Blue Cloud

**Associate Data Engineer - Rolling Meadows, IL** **Jun 2021 – Apr 2022**

- Built PySpark data pipelines in Azure Databricks for ETL from diverse file formats into Azure Data Lake Storage and Amazon S3, providing insights into customer usage patterns.
- Developed and deployed AWS Lambda functions in Python for data extraction, parsing, and ingestion of nested JSON from S3 into Snowflake.
- Engineered Snowflake Snowpipe for real-time data loading with internal and external stages, and defined roles, privileges, and virtual warehouse sizing for various workload types.
- Implemented databases, schemas, and tables in Databricks and Snowflake for diverse data handling and automated pipeline maintenance, improving ETL efficiency across multiple workloads.
- Provided L2/L3 production support for 100+ clients on the JLL Azara platform and contributed to integration infrastructure design across multiple teams.

George Mason University

**Graduate Teaching Assistant - Fairfax, VA** **Jan 2020 – May 2021**

- Assisted Computer Science Department in Graduate level courses – 'The Essentials of Computer Science' and 'Big data technologies'.
- Assisted the Information Sciences and Technology Department in Undergraduate level courses – 'Security Accreditation of Information Systems' and 'Information Security Principles'
- Collaborated with colleagues and professors to discuss student progress and helped develop skills during weekly office hours.

## Certifications

AWS Certified: Cloud Practitioner | 2023

Microsoft Certified: Azure Data Fundamentals | 2021

Microsoft Certified: Azure Data Engineer Associate | 2021

Snowflake Certified: SnowPro Core | 2021