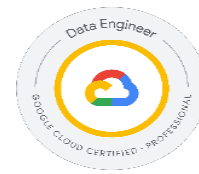**Surya Mithra Reddy Ram**
Senior AWS Data Engineer
Contact: (832)-786-8694
Email: rsmithrar@gmail.com | LINKEDIN

PROFESSIONAL SUMMARY

- Accomplished Senior Data Engineer with over 10 years of expertise in designing and managing data pipelines across diverse cloud platforms, including AWS, GCP, and Snowflake.
- Proficient in programming languages such as Python, PySpark, SQL, and Scala, enabling complex data transformations and real-time data processing.
- Extensive experience with data processing frameworks and tools like Spark, Dataproc, Google Cloud Dataflow, AWS EMR, and AWS Lambda, creating scalable and efficient data pipelines.
- Skilled in data warehousing techniques such as partitioning, clustering, and denormalization to optimize storage and query performance. Experience in integrating RESTful web services for seamless data exchange.
- Automation-driven mindset focused on improving maintainability, testability, and operational efficiency,
- Experienced in data visualization tools like Power BI and Advanced Excel, enabling effective storytelling and insights through interactive dashboards and reports.
- Proficient in ETL tools such as Informatica PowerCenter, Talend, AWS Glue, and SAP BODS to streamline data integration and transformation.
- Expertise in workflow orchestration tools like Apache Airflow and Google Cloud Composer to automate and monitor data pipelines for peak performance.
- Applied machine learning techniques for building classification and recommendation models using Python, supporting real-time, data-driven decision-making.
- Extensive experience with monitoring tools like AWS CloudWatch, AWS CloudTrail, and Google Stackdriver, ensuring real-time logging and pipeline reliability.
- Hands-on experience with version control systems, including Git, for effective code management, collaboration, and CI/CD integration.
- Skilled in building CI/CD pipelines using tools like Jenkins, Terraform, and Google Cloud Build for seamless deployment of production-ready data pipelines.
- Proficient in real-time data ingestion and processing using tools such as Google Pub/Sub, SLT, and HDFS, ensuring timely data availability for analytics and reporting.
- Familiar with regulatory compliance frameworks such as PHI, HL7, FHIR, and HIPAA, ensuring secure and compliant data handling.
- Adept at Agile methodologies, collaborating with cross-functional teams, and using tools like Jira for task management and sprint planning to deliver high-quality data solutions.

TECHNICAL SKILLS

Cloud Platforms	AWS, GCP, Azure, Snowflake, Databricks
Programming Languages	Python, PySpark, Scala, SQL, R, Node.js, React.
Data Processing	Spark, AWS Glue, EMR, AWS Lambda, Databricks, Hadoop, Hive, MapReduce, Talend, BODS
Data Warehouses	Redshift, Teradata, Oracle, SAP HANA, SAP ASE, MySQL, Amazon RDS, Google BigQuery, Cloud Spanner,.
ETL and Tools	Informatica PowerCenter, Pentaho, SSIS, SSRS, BODS, and SLT Replication, Talend, BODS.
Data Visualization	Power BI, Advanced Excel
Workflow Orchestration\ Security	Apache Airflow, Google Cloud Composer, AWS Step Functions, AWS Networking Fundamentals , VPC.
Version Control	GitHub.
CI/CD Tools	Jenkins, AWS CodeBuild, Cloud Build, Terraform

CERTIFICATIONS

- Certified Google Professional Data Engineer
- Certified **AWS** Solutions Architect Associate

EDUCATION

- Bachelor of Technology in Computer Science and Engineering, JNTU Anantapur, India

PROFESSIONAL EXPERIENCE

Client: Tennessee Farmers Insurance company, Nashville, TN   **Mar 2024 – Present**
Role: Senior Data Engineer
Responsibilities:
- Provided a comprehensive design and infrastructure solution for the migration project, ensuring all configurations and storage components were optimized for performance and scalability.
- Strong familiarity with AWS data and messaging services like Redshift, RDS, Kinesis, or Kafka was a plus, aiding in advanced analytics and high-throughput data streaming use cases.
- Created PySpark scripts to process large volumes of data, generating dynamic tables that facilitated advanced analytics and reporting capabilities for business stakeholders.
- A deep understanding of software engineering fundamentals was crucial, including version control with GitHub, collaborative coding practices, structured peer reviews, and managing feature branches effectively in a team environment.
- Developed and maintained scalable Java applications using core Java, Spring Framework, and Hibernate, ensuring high performance and responsiveness to requests from front-end users.
- Implemented object-oriented programming principles and design patterns to enhance code reusability and maintainability.
- Designed and implemented AWS Step Functions, triggering them based on events through Lambda, to serialize the user registration process and enhance user experience.
- Integrated AWS SDK for Java to enable seamless communication between Java applications and AWS services, enhancing data processing capabilities.
- Built data pipelines utilizing AWS Data Pipeline and AWS Glue to facilitate the loading of data from S3 into Redshift, ensuring timely and accurate data availability for analytics.
- Successfully migrated data from on-premise Oracle databases to Redshift using AWS Database Migration Service (DMS), ensuring data integrity and minimal downtime during the transition.
- Conducted data validation processes using Amazon Athena, leveraging S3 as the data source to ensure accuracy and consistency of migrated datasets.
- Established a data lake architecture using the Aurora database, enabling efficient data analysis and storage for various business intelligence applications.
- Transformed data according to specific business logic and loaded it into Redshift, providing stakeholders with reliable data for reporting and decision-making purposes.
- Managed IAM user creation and policy development, ensuring secure access controls and compliance with organizational security standards.
- Developed Lambda scripts to automate schema creation and historical data loads, streamlining the data ingestion process and reducing manual intervention.
- Created custom data pipelines to load data from S3 into DynamoDB tables using Python, enhancing data accessibility and performance for application developers.
- Designed and implemented QuickSight reports and dashboards to monitor daily job performance and data loads, providing insights into operational efficiency.
- Automated deployment processes using CI/CD pipelines via GitHub, CloudFormation, and CodePipeline, ensuring consistent and reliable infrastructure management.
- Developed Jenkins pipelines to manage resource creation and change sets, preventing duplicate resource creation and optimizing deployment workflows using Python.
- Designed and implemented Lakehouse architectures using Delta Lake, ensuring ACID compliance, schema enforcement, and time travel for efficient big data processing and real-time analytics.
- Optimized Spark jobs through adaptive query execution, caching, partitioning, and broadcast joins, achieving a 30-50% reduction in processing time and improved cluster utilization.
- Implemented Delta Lake optimization techniques, including Z-Ordering, data skipping, and Change Data Capture (CDC), to enhance query performance and maintain data consistency.
- Built real-time data pipelines using Spark Structured Streaming and Autoloader, ingesting high-velocity event data from

Kafka, Kinesis, and IoT devices for near-instantaneous insights.

Environment/Tools: AWS, EC2, S3, Glue, Java, Delta Lake, Node.js, Databricks Workflows, Databricks Jobs, Data Pipeline, DMS, Lambda, Dynamo DB, JavaServer Faces , Aurora, Athena, Cloud Formation, Code Pipeline, Python, Pyspark, Airflow, PostgreSQL, Oracle, Tableau, Unix, Shell Script, GitHub, oracle data modeler.

Client: Starbucks, Seattle, WA **Aug 2021 – Feb 2024**
Role: AWS Data Engineer
Responsibilities:
• Built scalable ELT pipelines to ingest CRM data into Snowflake using FiveTran, transforming OLTP data to OLAP format for Redshift storage and analytics.
• Captured real-time IoT data via Amazon Kinesis and automated pipelines for timely ingestion and processing with Airflow orchestration.
• Architected and developed microservices using Java and Spring Boot, deployed on AWS ECS (Elastic Container Service) for improved scalability and fault tolerance.
• Designed and managed data orchestration workflows using native AWS services to enable seamless micro front-end and backend integration for high-performance, distributed cloud applications.
• Built and maintained secure, cloud-optimized APIs using AWS API Gateway to support reliable, scalable communication and integration with data services and external systems.
• Developed data access layers using JPA and Hibernate to interact with AWS DynamoDB and RDS, ensuring data integrity and security.
• A deep understanding of software engineering fundamentals was crucial, including version control with GitHub, collaborative coding practices, structured peer reviews, and managing feature branches effectively in a team environment.
• Developed and maintained scalable Java applications using core Java, Spring Framework, and Hibernate, ensuring high performance and responsiveness to requests from front-end users.
• Designed staging processes in Amazon S3 for real-time data, applying scalable transformations using EMR and Python.
• Developed AWS Glue pipelines to process real-time data into Redshift and DynamoDB, employing advanced partitioning and clustering to enhance analytics capabilities.
• Developed and optimized Databricks notebooks for large-scale data transformations in a cloud environment.
• Orchestrated ETL pipelines across platforms using AWS Step Functions, integrating data with Python and Spark transformations.
• Applied machine learning models like classification and recommendation systems to real-time datasets using Python and Scala for advanced predictive analytics.
• Automated ETL pipelines with Apache Airflow, enhanced scheduling with cron jobs, and monitored workflows with comprehensive logging mechanisms.
• Clear and concise technical documentation skills were important for maintaining codebases, onboarding new team members, and supporting long-term maintenance and scalability of solutions.
• Created Python DAGs in Airflow to orchestrate end-to-end data pipelines and integrated real-time datasets via EMR.
• Streamlined CI/CD processes using Jenkins, ensuring automated deployment of data pipelines to production.
• Version-controlled Python and Scala scripts in Git, enabling smooth collaboration and clean versioning within Agile teams.

Tech Stack: Snowflake, Amazon Kinesis, Amazon S3, AWS Step Functions, AWS Glue, EMR, Redshift, DynamoDB, SQL, Python, Machine Learning, Apache Airflow, Unity Catalog, Databricks Workflows, Databricks Jobs, Cron Jobs, Python DAGs, Git, Jenkins, CI/CD Pipelines, Agile Methodology.

Client: CVS, Woonsocket, RI **Jun 2020- Aug 2021**
Role: AWS Data Engineer
Responsibilities:
• Migrated healthcare data from Teradata and Oracle to Amazon Redshift using Talend and AWS Glue while ensuring compliance with HL7 and FHIR standards.
• Designed data mirroring techniques, staging PHI-compliant healthcare data securely in Amazon S3.
• Developed and maintained scalable Java applications using core Java, Spring Framework, and Hibernate, ensuring high performance and responsiveness to requests from front-end users.
• Designed staging processes in Amazon S3 for real-time data, applying scalable transformations using EMR and Python.
• Developed AWS Glue pipelines to process real-time data into Redshift and DynamoDB, employing advanced partitioning and clustering to enhance analytics capabilities.
• Developed Python and Spark scripts on EMR and AWS Lambda to transform healthcare data for analytics in Redshift.

- Created and maintained table schemas in Redshift, ensuring seamless integration with downstream systems.
- Captured and processed real-time healthcare data via APIs, storing it in efficient formats for scalability and analytics.
- Extracted Salesforce CRM data with AWS Glue, transforming and loading into Redshift to support business needs.
- Built ELT pipelines with audit tracking to monitor and log data migrations and transformations for compliance purposes.
- Automated batch processing with AWS Step Functions, AWS Glue, EMR, and AWS Lambda to handle large volumes of healthcare data.
- Automated batch processing with Databricks Jobs, reducing operational overhead and manual interventions.
- Monitored pipelines using Amazon CloudWatch and AWS CloudTrail, ensuring real-time tracking of healthcare data workflows.
- Collaborated with cross-functional teams in an Agile environment to deliver high-quality Java applications on AWS, participating in sprint planning, code reviews, and retrospectives.
- Mentored junior developers in Java programming and AWS best practices, fostering a culture of continuous learning and improvement.
- Managed metadata using Amazon RDS for tracking schema versions, data sources, and transformation logic

Tech Stack: Talend, AWS Glue, JavaServer Faces , Amazon Redshift, Teradata, Oracle, HL7 & FHIR Standards, Amazon S3, PHI Compliance, Python, Apache Spark, Amazon EMR, AWS Lambda, Salesforce CRM, Java, AWS Step Functions, Amazon RDS, Git, Jira, Agile Methodology.

Client: Nextera, Juno Beach, FL   **Sep 2018 – May 2020**
Role: Data Engineer
Responsibilities:
- Built a framework to accommodate Full loads & Incremental loads across AWS and GCP environments.
- Worked on Data Analysis, validations, and audit framework implementation.
- Responsible for collecting client requirements for specific data sources and designing solutions accordingly.
- Developed & implemented solutions using Apache Spark and Python API to load data from AWS S3 and Google Cloud Storage (GCS).
- Created dimension and fact tables and loaded transformed data into AWS Redshift and Google BigQuery.
- Applied end-to-end unit testing and documented results for each requirement story, reviewing them with the test lead before production deployment.
- Worked in Agile methodology to meet deadlines for full ELT cycle requirements.
- Collaborated closely with Business users, interacting with ETL developers, Project Managers, and QA teams.
- Created different KPIs using calculated key figures and parameters.
- Automated manual processes using Python scripts to improve efficiency and save time.
- Optimized performance and processing for existing data flows.
- Responsible for the documentation, design, development, and architecture of visualization reports.
- Handled the installation, configuration, and support of a multi-node setup for AWS EMR and GCP Dataproc.
- Developed automation solutions using Python to streamline processes across AWS and GCP environments.

Tech Stack: AWS S3, AWS Redshift, Google Cloud Storage (GCS), Google BigQuery, Apache Spark, Agile Methodology, Fact and Dimension Tables, Python Automation, KPIs, AWS EMR, Google Cloud Dataproc.

Client: Indiumsoft, Hyderabad, India
Role: Big Data Engineer   **Aug 2017– Sep 2018**
Responsibilities:
- Built and maintained ETL pipelines using Informatica PowerCenter to integrate sales data into Teradata while applying SCD Type 1 and Type 2 for handling historical data.
- Designed and optimized Informatica mappings for large-scale data loads using FastLoad and MultiLoad, with staging tables in Oracle DB to streamline transformations.
- Developed star schema-based data marts in Teradata, including fact and dimension tables, to support advanced reporting and analytics.
- Created ETL workflows to load data into data marts, enabling advanced business insights and KPI reporting.
- Integrated KPI logic into ETL workflows to ensure accurate business metric generation.
- Conducted extensive testing, including unit, integration, and user acceptance testing, to ensure data reliability.
- Documented ETL processes, transformations, and mappings, creating user manuals and data dictionaries for maintenance and user reference.

Tech Stack: Informatica PowerCenter, Teradata, Oracle DB, FastLoad, MultiLoad, Star Schema

Client: Spurtree Technologies Inc, Bangalore, India   **Jun 2016 – Aug 2017**
Role: ETL Developer
Responsibilities:

- Designed and implemented customized database systems to meet client requirements.
- Authored comprehensive design specifications and documentation for database projects.
- Developed ETL pipelines using Pentaho for seamless integration of data from various sources.
- Troubleshot and optimized MapReduce jobs to resolve failures and improve performance.
- Facilitated data import/export across multiple systems and the Hadoop Distributed File System (HDFS).
- Built scalable and distributed data solutions utilizing Hadoop, Hive, MapReduce, and Spark.
- Transformed structured and semi-structured data using tools like Hive and Spark.
- Created detailed user documentation for Hadoop ecosystems and processes. Executed Hive queries to perform in-depth data analysis and validation.

Tech Stack: MapReduce, Pig, Hive, Hadoop, Cloudera, HBase, Sqoop