**Roopalatha Agarala**

**Data Engineer**

saisri@mastroservices.com

+1 (469) 846-8345

**PROFESSIONAL SUMMARY:**

* **Data Engineer** with 9 years of experience in designing and implementing production-grade data warehousing and data lake solutions leveraging large-scale data technologies on **Amazon Web Services (AWS)** and **MS Azure** cloud platforms.
* Over 7 years of expertise in complex **cloud applications**, with more than 5 years of hands-on experience with **AWS services** including **RDS, S3, Step Functions, Glue, Lambda, Athena, IAM, CodePipeline, CloudFormation, SNS**, and **SQS**.
* Experienced in diverse domains such as **Healthcare, Finance, e-commerce,** and **Oil & Gas**. Proficient in **Python** libraries like **Pandas** and **PySpark** for data analysis, cleaning, visualization, and model building.
* Experience in Data Warehousing, building **ETL pipelines.**
* Experience with **Storm** and **Kafka** for the real time processing of data.
* Efficient in all phases of the development lifecycle, coherent with Data Cleansing, Data Conversion, Data Profiling, Data Mapping, Performance Tuning, and System Testing.
* Experience with **Amazon Web Services** (Amazon EC2, Amazon S3, Amazon RDS, Amazon Elastic Load Balancing, Amazon SQS, AWS Identity and access management, Amazon kinesis, AWS Cloud Watch, Amazon EBS, Amazon CloudFront, VPC, DynamoDB, Lambda and Redshift).
* Utilized **Databricks** for large-scale data processing and analysis, optimizing Spark jobs for performance and cost efficiency.
* Experienced in using **streaming pipelines** that ingest data from a messaging system such as **Apache Kafka**, process the data in real-time using **Apache Spark**, and then store the results in a data store such as Databricks.
* Experienced in Developing Spark applications using RDD, **Spark SQL**, **Data Frames** and **Spark Streaming**, Spark-SQL in Databricks for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
* Defined policies using **AWS IAM** that specified the actions that users can perform on AWS resources.
* Good experience on Data Modelling (Dimensional and Relational) concepts like **Star-Schema** Modeling, **Snowflake Schema** Modeling, and **Fact** and **Dimension** Tables.
* Experience with **snowflake multi-cluster** warehouses.
* Expertise in **Terraform** for multi cloud deployment using single configuration.
* Experience in building **Snow pipes** and using **snowflake clones** and **time travel.**
* Basic exposure to TypeScript, API development, GraphQL, NPM, and Docker, enhancing versatility in modern development practices.
* Hands-on experience on **MS** **Azure** **Cloud suite:** Azure SQL Database, Azure Data Lake (ADL), Azure Data Factory (ADF), Azure SQL Data Warehouse, Azure Service Bus, DynamoDB, Azure Key Vault, Azure Analysis Service (AAS), Azure Blob Storage, Azure Search, Azure App Service, and Azure Data Platform Services.
* Experienced in using **Git, Bitbucket,** and **Jenkins** for version control and continuous integration, ensuring high-quality, reliable, and reproducible data pipelines and workflows.
* Hands-on experience in visualizing the data using **Power BI, Tableau, Python** (Pandas, matplotlib, NumPy, SciPy).
* Expertise in implementing and managing robust and scalable **SQL** and **NoSQL** databases such as **MySQL**, **MongoDB**, and **SQL servers**, **PostgreSQL.**
* Experienced with **Agile Methods,** with a **Python**-based environment, Scrum stories, and sprints, as well as data analytics, data wrangling, and Excel data extraction.
* Strong experience in working with **UNIX/LINUX** environments.
* Excellent communication and interpersonal skills with ability in resolving complex business problems.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **Category** | **Skills** |
| Cloud Platforms | AWS (RDS, S3, Glue, Lambda, Athena, Redshift), Azure (SQL Database, Data Lake), Terraform |
| Data Engineering | ETL, Data Warehousing, Data Lakes, Real-time Processing (Kafka, Storm), Data Modeling, Snowflake, Data Governance, Security Compliance, Disaster Recovery |
| Programming | Python (Pandas, PySpark), Scala, SQL |
| Big Data Technologies | Databricks, PySpark, Apache Kafka, Spark SQL, AWS Glue, AWS Kinesis, AWS Athena, Apache Airflow |
| Database Management | MySQL, MongoDB, SQL Server, PostgreSQL, Azure Synapse Analytics, Azure SQL Database, Cassandra DB |
| DevOps/CI/CD | Git, Bitbucket, Jenkins, CodePipeline, CloudFormation, Continuous Integration, Continuous Deployment, Terraform |
| Visualization | Power BI, Tableau, Matplotlib, NumPy, SciPy, Quick Sight |
| Others | Kafka API, GraphQL, NPM, Docker, Azure Active Directory, Agile Methodologies, UNIX/LINUX Environments |

**PROFESSIONAL EXPERIENCE**

**Client: LabCorp, Durham, NC June 2021 – Till Now**

**Role: Sr Data Engineer**

**Responsibilities:**

* Worked closely with users and business analysts to gather requirements and designed appropriate data solutions to fulfill the business needs effectively.
* Created an inbound and outbound framework to optimize data loading (ETL) into Databricks from various sources, as well as loading data from Databricks to SAP HANA and Snowflake.
* Designed and implemented scalable and highly available cloud architectures on AWS.
* Migrated on-premises data infrastructure to AWS cloud, ensuring minimal downtime and data integrity.
* Designed and implemented scalable data lakes on Amazon S3 for storing structured and unstructured data.
* Implemented Kafka streaming for replicating real-time data from SQL server for on-demand orders. Utilized medallion architecture to transform and store data into bronze, silver, and gold tables, and loaded final output to SAP HANA for reporting purposes.
* Integrated CodePipeline with source control systems such as GitHub, Bitbucket, and AWS CodeCommit.
* Designed and implemented relational databases using Amazon RDS for scalable and secure data storage.
* Performed database performance tuning and optimization for complex queries and transactions.
* Worked on creating data pipelines and transforming data to a structure that is relevant to the problem by selecting appropriate techniques.
* Integrated CloudFormation with CI/CD pipelines for automated infrastructure provisioning.
* Architected and deployed robust ETL pipelines using AWS Glue, Lambda, and Step Functions to automate data ingestion, transformation, and loading processes, enhancing data workflow efficiency.
* Developed and maintained Glue scripts in Python and Scala for complex data transformations.
* Automated data processing and ETL workflows using Step Functions to coordinate multiple AWS services.
* Ensured data security and compliance by configuring and managing AWS IAM roles and policies, implementing encryption using AWS KMS, and monitoring data access with AWS CloudTrail.
* Implemented S3 bucket policies, ACLs, and encryption to ensure data security and compliance.
* Integrated S3 with AWS analytics and machine learning services such as Glue, Athena, and SageMaker.
* Worked with building data warehouse structures, and creating facts, dimensions, and aggregate tables, by dimensional modeling, and Star and Snowflake schemas.
* Implemented Snowflake data sharing to seamlessly share data with the client's Snowflake account.
* Performed ETL operations for Data cleansing, Filtering, Standardizing, Mapping, and Transforming of extracted data from multiple sources such as AWS, SQL Server, DynamoDB, Oracle, My SQL DB, and PostgreSQL.
* Developed and maintained SQL scripts for data manipulation, aggregation, and reporting purposes in alignment with healthcare data standards.
* Developed data integration solutions using AWS SNS and SQS for reliable message delivery and decoupling of services, improving system scalability and fault tolerance.
* Configured Spark streaming to receive real time data from Kafka and store the steam data to HDFS using Scala.
* Worked with Terraform Templates to automate the Azure IA as virtual machines using terraform modules and deployed virtual machine scale sets in production environments.
* Implemented data clean rooms in Snowflake and Databricks to conduct comprehensive analysis by combining client data securely.
* Created Databricks notebooks using PySpark, and Spark SQL to read and write JSON, DynamoDB, CSV, and Parquet.
* Created workflows within Databricks to automate daily, weekly, and monthly data loads.
* Utilized Databricks for large-scale data processing and analysis, optimizing Spark jobs for performance and cost efficiency.
* Performed CI/CD to Dev, QA and PROD environments using bitbucket and Jenkins.

**Environment:** Databricks, Kafka, PySpark, Terraform, AWS RDS, AWS S3, Athena, Snowflake, DynamoDB, Jira, AWS Kinesis, AWS IAM, AWS Glue crawlers, Glue jobs, CodePipeline, Lambda functions, Glue triggers, SQL, AWS Step Functions, AWS Redshift, AWS CloudWatch, CloudFormation, SQS, SNS, Agile-Scrum, SQL Server, Medallion architecture, AWS CodeCommit, Star and Snowflake schemas, Oracle, PostgreSQL, Healthcare data standards, Spark streaming with Scala, Azure IA (Infrastructure as virtual machines), Data clean rooms, Bitbucket, Jenkins

**Client: ExxonMobil, Houston, TX Apr 2019 – May 2021**

**Role: Data Engineer**

**Responsibilities:**

* Designed and implemented end-to-end data pipelines on AWS and Databricks platforms to ingest, process, and analyze large volumes of data for various business use cases.
* Utilized AWS Glue for data cataloging, ETL processing, and schema evolution, ensuring data consistency and availability across the organization.
* Implemented scalable data lake architectures on AWS S3 and Delta Lake, optimizing storage and query performance for efficient data access and analysis.
* Protype done with HDP Kafka and Storm for click stream application.
* Integrated Glue with other AWS services such as S3, RDS, Redshift, and Athena for seamless data processing.
* Developed and maintained PySpark applications for data processing, transformation, and machine learning tasks, ensuring high-performance and fault-tolerant data processing.
* Migrated on-premises databases to Amazon RDS with minimal downtime using AWS Database Migration Service.
* Enforced data security and compliance using AWS IAM roles, encryption, and parameter groups.
* Performed bulk loading from the external stage (AWS S3), internal stage to snowflake cloud using the COPY command.
* Designed and implemented complex workflows and state machines using AWS Step Functions.
* Implemented Lambda to configure DynamoDB Auto Scaling feature and implemented Data Access Layer to access AWS DynamoDB data.
* Conducted regular audits and reviews of CloudFormation templates for optimization and best practices.
* Implemented event-driven architectures by integrating Lambda with AWS services such as S3, SNS, SQS, and DynamoDB.
* Loaded data into snowflake tables from the internal stage using Snow SQL.
* Written complex Snow SQL scripts in snowflake cloud data warehouse to business analysis and reporting.
* Written Kafka API to collect events from the front end.
* Monitored and managed pipeline execution and performance using AWS CloudWatch and CodePipeline Console.
* Ensured data security and compliance by configuring and managing AWS IAM roles and policies, implementing encryption using AWS KMS, and monitoring data access with AWS CloudTrail.
* Used SNOW PIPE for continuous data ingestion from the S3 bucket.
* Used Terraform in managing resource scheduling, disposable environments, and multi tier applications.
* Created clone objects to maintain zero-copy cloning.
* Collaborated with cross-functional teams to understand business requirements and translate them into technical solutions, driving innovation and business value through data-driven insights.
* Implemented data governance and security best practices on AWS and Databricks platforms, ensuring compliance with regulatory standards and data privacy policies.

**Environment:** Snowflake, AWS RDS, AWS Data Pipeline, Athena, Jira, Apache Airflow, CodePipeline, PySpark, AWS Glue, Glue jobs, Lambda functions, IAM, Glue triggers, AWS Step Functions, CloudFormation, SQS, SNS, AWS Dynamo, Kafka, AWS S3, Agile-Scrum, HDP, Kafka API, Snow SQL, Snow PIPE, Terraform, Zero-copy cloning

**Client: Credit Suisse, Raleigh, NC. Mar 2017 – Apr 2019**

**Role: Data Engineer**

**Responsibilities:**

* As a data engineer, I am responsible for processing payment-related data such as credit card information, bank account details, and transaction history. It is important to ensure that this data is protected and kept confidential as per the guidelines set forth in relevant compliance regulations.
* I have ensured that the e-commerce platform is able to implement this mechanism accurately and in compliance with the requirements of relevant consumer protection regulations.
* Extracted data from Azure Blob Storage and performed necessary Transformations and Aggregation in the pipelines to build the data model and stored the processed data in Azure Synapse Analytics (formerly known as Azure SQL Data Warehouse).
* Processed the data using data pipelines using Azure Data Factory and wrote PySpark scripts inside the pipelines to clean and process the data.
* Processed data from different files (CSV, JSON, AVRO, and Parquet) by writing Python scripts in Spark, then loaded the processed data to Azure Synapse Analytics.
* After Data processing, we connect Azure Synapse Analytics to Power BI for creating dashboards and report generation.
* Demonstrated Key Performance Indicator (KPI) dashboards using Power BI.
* Responsible for data transfer between different Azure computing and storage services using Azure Data Factory.
* Developed custom ETL solutions, Batch Processing for data ingestion, and pipelines to move data in and out of Azure Databases using Python.
* Developed ETL processes in Azure Databricks to relocate processed data from external Databases into Azure SQL Database.
* Written Storm topology to accept the events from Kafka producer and emit into Cassandra DB.
* Automate Datadog Dashboards with the stack through Terraform Scripts.
* Worked with Azure Active Directory (AAD) to set up user roles with corresponding user and group policies using JSON and add project users to the Azure account with multi-factor authentication enabled and least privilege permissions.
* Used Azure Data Lake Analytics to query large amounts of data available in Azure Data Lake Storage.
* Designed and implemented data pipelines and ETL processes to move data from various sources into Azure Synapse Analytics for analysis and reporting purposes.
* Worked in agile development environments for Project Management and delivery.
* Adapted to changing requirements and priorities through agile methodologies.
* Worked with data stakeholders to prioritize and manage backlogs and user stories.

**Environment:** Spark, python, MySQL, PySpark, Azure synapse, Quick Sight, power bi, Azure Blob Storage, Azure Data Factory, Azure Databricks, Storm, Kafka, Cassandra DB, Datadog, Terraform, Azure Active Directory, Azure Data Lake Analytics, Agile methodologies

**Client: Apollo, India May 2015 - Nov 2016**

**Role: SQL Developer**

**Responsibilities:**

* Designed and developed complex SQL queries to extract and analyze data from databases.
* Designed and implemented database schemas, tables, views, and indexes.
* Developed and maintained stored procedures, functions, and triggers.
* Created ETL processes to extract, transform, and load data from different data sources by using command line interface (CLI).
* Performed database tuning and optimization to improve query performance.
* Conducted database backups and disaster recovery tests.
* Designed and optimized database schemas for performance and scalability by using DBMS.
* Implemented security policies and access controls to ensure data integrity and confidentiality.
* Collaborated with cross-functional teams to develop and implement data-driven solutions.
* Trained and mentored junior SQL developers on best practices and techniques.
* Designed and developed SQL Server Database, Tables, Indexes, Stored procedures, Views, and Triggers.
* Developed stored procedure, lookup, execute the pipeline, data flow, copy data.

**Environment:** SQL, CLI, DBMS (Database Management Systems), Data integration, Data transformation, Data Migration, Data Warehousing.

* **EDUCATION:**Bachelors of Technology, Electronics and communication Engineering in JNTU, Anantapur - June 2011 to May 2015.